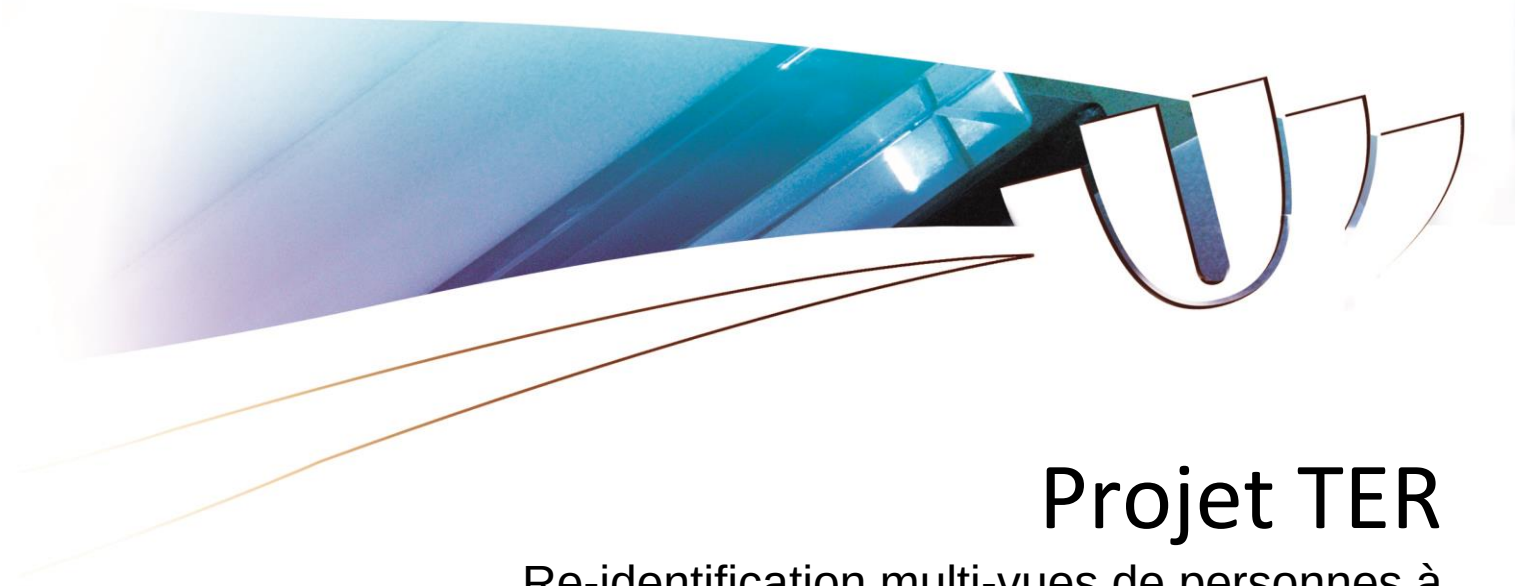

Projet TER

Re-identification multi-vues de personnes à l'aide d'apprentissage profond

Réalisé par :

Lafdhal Ahmed

Encadrée par :

Itheri Yahiaoui

UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

Table des matières

UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

I. Introduction

La re-identification des personnes en multivue est un domaine de recherche en pleine expansion dans le domaine de computer vision Elle consiste à identifier une personne à partir de plusieurs captures vidéo provenant de différentes caméras. Cette tâche est particulièrement difficile en raison de la variation des conditions d'éclairage, de l'angle de vue et de la qualité de l'image dans les différentes captures vidéo.

Prenons par exemple un cas où une personne est suivie par une caméra dans un centre commercial et quitte la zone de surveillance de cette caméra pour passer sous la surveillance d'une autre caméra. L'identification de cette personne à partir de ces deux captures vidéo est une tâche complexe, car les deux captures vidéo peuvent avoir des différences significatives en termes de qualité d'image, d'angle de vue et de conditions d'éclairage.
Je suis convaincu que l'importance de la re-identification des personnes en multivue réside dans son utilisation pour la surveillance et la sécurité. En effet, cette technologie permet de suivre les déplacements des personnes dans un environnement donné et d'identifier les personnes impliquées dans des activités illégales ou suspectes.

Je me suis intéressé à ce domaine et ai décidé de travailler sur ce sujet visant à proposer une solution pour la re-identification des personnes en multivue en utilisant les techniques de deep learning.
Mon objectif est d'approfondir ma compréhension de ce domaine en examinant les travaux antérieurs des chercheurs et en étudiant les différents modèles de deep learning utilisés pour la re-identification des personnes en multivue. Je prévois de tester ces modèles sur des datasets appropriés pour évaluer leur précision et leur efficacité dans la résolution de cette tâche complexe.
Dans ce rapport, je vais d'abord présenter les travaux antérieurs dans le domaine de la re-identification des personnes en multivue et je présenterai les résultats.

II. Défis de la réidentification des personnes en multivue

La re-identification des personnes en multivue est une tâche complexe et présente plusieurs défis importants. Avant même de pouvoir effectuer la Re-ID, le système doit détecter une personne et définir la boîte englobante de la personne dans une image. Cependant, le corps humain est très déformable, ce qui rend la détection d'objets déformables en soi un défi important.
Les changements d'éclairage sont un autre défi important. Les variations d'intensité de la lumière naturelle, l'ombre, la lumière réfléchie par des surfaces colorées, ainsi que l'éclairage intérieur peuvent causer des variations dans les couleurs et les ombres des personnes, même sur des caméras différentes.
La faible résolution des caméras est un autre obstacle important. Les anciens systèmes de vidéosurveillance peuvent avoir des caméras de faible résolution, ce qui rend la reconnaissance des personnes encore plus difficile en raison du manque d'informations.
L'occlusion partielle ou complète des personnes par d'autres personnes est également un défi important en environnements encombrés. Les algorithmes de Re-ID doivent pouvoir extraire des caractéristiques de

personnes même lorsqu'elles sont partiellement cachées.
Les vêtements uniformes dans les écoles et certains lieux de travail peuvent également compliquer les choses en confondant les algorithmes de Re-ID qui extraient des informations à partir de l'apparence des personnes.
Enfin, la mise à l'échelle est un défi important. Les espaces publics sont couverts par des milliers de caméras et les technologies actuelles ne sont qu'à leurs débuts pour résoudre le problème de la surveillance multi-caméras.

III. Datasets

Dans le domaine de la réidentification des personnes et comme dans tous les domaines de deep learning et computer vision en particulier, il est important d'avoir des ensembles de données volumineux et annotés pour entraîner et évaluer les algorithmes. Dans le cas de la réidentification des personnes, plusieurs ensembles de données ont été créés à cet effet, tels que Market-1501, DukeMTMC-reID, CUHK03, MSMT17 et Mars. Pour la Re-ID de personne, les ensembles de données utilisés peuvent être généralement divisés en deux catégories : les ensembles de données basés sur des images et les ensembles de données basés sur des vidéos. Toutefois, les problèmes tels que l'occultation, les changements d'éclairage, les changements de vue de caméra, les changements de posture et les vêtements similaires qui ne peuvent toujours pas être résolus de manière satisfaisante, rendent de nombreux algorithmes inapplicables dans des situations réelles.

A. Datasets pour Reid basé sur l'Image

- Market-1501 : est l'un des ensembles de données les plus populaires, contenant plus de 32 000 images de 1 501 personnes prises par six caméras différentes.
- DukeMTMC-reID : est un autre ensemble de données largement utilisé, avec plus de 36 000 images de 1 404 personnes capturées par huit caméras différentes.
- CUHK03 : est un ensemble de données plus petit, mais très populaire, contenant plus de 14 000 images de 1 467 personnes capturées par deux caméras différentes.
- MSMT17 : est un ensemble de données plus récent et plus grand, contenant plus de 126 000 images de 4 101 personnes prises par 15 caméras différentes.

B. Datasets pour Reid basé sur Vidéo

- MARS : Il s'agit d'un ensemble de données de Re-ID de personnes à grande échelle basé sur des vidéos, qui contient plus d'un million de cadres extraits de plus de 1 200 séquences vidéo. Les vidéos ont été capturées dans des environnements réels, avec des scénarios tels que des foules, des couloirs et des zones extérieures. Les sujets portent une variété de vêtements, ce qui rend le défi de la Re-ID encore plus difficile.
- PRID 2011 : Il s'agit d'un ensemble de données de Re-ID de personnes basé sur des vidéos, qui contient deux séquences vidéo capturées à partir de caméras opposées dans une zone piétonne. L'ensemble de données contient des défis tels que des changements de point de vue, des

occlusions et des changements de luminosité.

- iLIDS-VID : Il s'agit d'un ensemble de données de Re-ID de personnes basé sur des vidéos, qui contient des séquences vidéo capturées à partir de deux caméras dans un couloir. L'ensemble de données contient des défis tels que des changements de pose, des changements d'expression faciale et des changements de vêtements, ainsi que des changements de luminosité.

IV. Méthode d'évaluation :

Les méthodes d'évaluations les plus utilisé dans la réidentification des personnes sont les courbes CMC (Cumulative Matching Characteristics) et la MAP(mean average precision)

- CMC : Les courbes cmc consiste à calculer qu'une correspondance correcte apparaisse parmi les 5 meilleurs résultats.
  L'accuracy de la re identification d'une image dans le top-5 est calculé de la manière suivante :

$$Acc_k^i = \begin{cases} 1, & \text{if the top-}k\text{ ranked gallery samples} \\ & \text{contain the sample(s) of query } i; \\ 0, & \text{otherwise.} \end{cases}$$

Et donc pour les N images

$$\text{CMC-}k = \frac{1}{N}\sum_{i=1}^{N} Acc_k^i$$

- MAP : le mean average precision mesure la précision moyenne de l'algorithme de Re-ID pour identifier les bonnes correspondances dans les différentes vues d'un ensemble de données.

V. Deep Learning

A. Les fonctions de perte couramment utilisées

1. Classification (ID) et Vérification Loss:

Dans cette méthode chaque personne est considérée comme une classe différente et l'ID du piéton est utilisé comme étiquette de classification pour entraîner un classifieur basé sur l'apprentissage profond.
Cette méthode utilise une couche fully connected (FC) pour la classification et la fonction

d'activation soft max pour transformer les vecteurs de caractéristiques en une distribution de probabilités sur l'ensemble des identités possibles
L'ID loss est largement utilisée car elle permet un entraînement facile et l'identification de cas difficiles.
Voici la cross entropy utilisé pour la tâche de la classification multiclasse:

$$\mathcal{L}_{id} = -\sum_{a=1}^{K} q\left(x_a\right) \log p\left(y_a | x_a\right)$$

La vérification loss est utilisée pour valider le modèle ou autrement dit guider le modèle contrairement à la classification loss il prend en entrée deux images et mesure la similarité entre ces deux images et donne en sortie si oui ou non ces deux images se correspondent la cross entropy utilisé et la suivante :

$$\mathcal{L}_v = -y_{ab} \log p\left(y_{ab} | f_{ab}\right) - (1 - y_{ab}) \log \left(1 - p(y_{ab} | f_{ab})\right)$$

Les chercheurs ont opté pour la combinaison de ces deux loss pour avoir un meilleur résultat donc le loss hybride est la somme de ces deux loss

L = L(id)+L(v)

2. Contrastive loss

L'objectif de la Contrastive Loss est de minimiser la distance entre les paires d'images appartenant à la même personne (images similaires), tout en maximisant la distance entre les paires d'images appartenant à des personnes différentes (images dissemblables). Avec cette fonction de cout le modèle apprendra à représenter les images de personnes similaires dans la même région, et à séparer les images de personnes différentes.
Plus précisément, si nous avons deux images d'une même personne, nous voulons que la distance entre leurs représentations dans l'espace de caractéristiques soit le plus proche de 0. Si nous avons deux images de personnes différentes, nous voulons que la distance entre leurs représentations dans l'espace de caractéristiques soit grande.
Cette fonction de coût a été appelée contraste car elle met en contraste les images de personnes similaires et dissemblables.

3. Triplet Loss

Le triplet loss est la fonction de cout la plus utilisé dans le domaine de la Reid des personnes.
Le triplet loss utilise des triplets d'images pour apprendre un modèle qui peut distinguer les caractéristiques physiques de différentes personnes. Chaque triplet est composé de trois images : une image fixe, une image positive et une image négative. L'image fixe est une image d'une personne donnée, la positive est une image de la même personne dans l'image fixée, et la négative est une image d'une personne différente que celle de l'image fixe.
Le but du triplet loss est d'apprendre un modèle qui peut distinguer l'image fixe des images positives et négatives. Pour ce faire, il compare les distances entre les caractéristiques physiques des trois images du triplet. La distance entre l'image fixe et la positive est appelée distance intra-classe, tandis que la distance entre l'image fixée et la négative est appelée distance inter-classe.
L'objectif est de minimiser la distance intra-classe et de maximiser la distance inter-classe.
Plus précisément, le triplet loss est défini comme suit : pour chaque triplet (a, p, n), la perte est donnée par max (0, d(a, p) - d(a, n) + m), où d(a, p) est la distance entre l'image fixée et l'image positive, d(a, n) est la distance entre l'image fixée et l'image négative, et m est une marge qui permet de définir une distance minimale entre les distances intra-classe et inter-classe.
En minimisant cette fonction de perte, le modèle apprend à distinguer les caractéristiques physiques des différentes personnes, ce qui permet de reconnaître de manière fiable les individus à partir de leurs images.
Les chercheurs ont montré qu'en le triplet loss et la classification loss le modèle arrive à mieux réidentifier les personnes

B. Modèles de Re-Id

J'ai choisi de présenter le modèle omniscale (OSNet) qui a atteint l'état de l'art dans le domaine de la réidentification des personnes en multi vue

1. Détail du modèle

Dans l'article les auteurs proposent une architecture de réseau de neurones convolutionnels pour la réidentification de personnes. Leur méthode se compose de deux étapes principales : la première étape consiste à extraire des caractéristiques visuelles à différentes échelles à l'aide de blocs de convolution à différentes résolutions. La deuxième étape consiste à fusionner les caractéristiques omniscales extraites par les blocs de convolution pour obtenir une représentation finale qui est utilisée pour la réidentification de personnes. Cette méthode permet d'apprendre des représentations de haute qualité pour la ré-identification de personnes à différentes échelles, ce qui améliore les performances des systèmes de ré-identification de personnes.
Les auteurs de cet article ont opté pour le depthwise séparable convolutions afin de réduire le nombre de paramètres.
Le depthwise séparable convolution est une technique de convolution en deux étapes qui

permet de réduire le nombre de paramètres et la complexité des modèles de réseaux de neurones tout en améliorant leur précision.
Dans la méthode classique la première étape est appelée depthwise qui applique une convolution séparée à chaque canal d'entrée et la deuxième méthode appelée pointhwise qui consiste à combiner les sorties de chaque canal d'entrée à travers une convolution classique de taille 1*1.
Mais dans leur article ils ont trouvé que (pointwise → depthwise au lieu de depthwise → pointwise) était plus efficace.
Ils ont appelé cette couche Lite

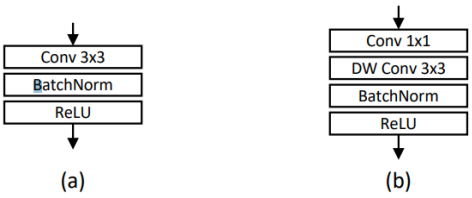

Fig. 3. (a) Standard and (b) Lite 3×3 convolution. DW: Depth-Wise.

Le bloc principal de l'architecture est le bottleneck résiduel, équipé de la couche Lite 3 x 3.
Pour avoir un apprentissage à plusieurs échelles, les auteurs ont étendu la fonction résiduelle F qui est une couche Lite 3x3 qui apprend sur une seule échelle en ajoutant une nouvelle dimension,
L'exposant t, qui représente l'échelle. Pour F^t, avec t > 1
Ce qui nous donne la fonction d'apprentissage suivante

$$\tilde{\boldsymbol{x}} = \sum_{t=1}^{T} F^t(\boldsymbol{x}), \quad \text{s.t.} \quad T \geqslant 1.$$

Ils ont opté aussi une AG (agregation gate) pour combiner dynamiquement les différents features donc les différents poids au lieu d'être combiner une fois à la fin de l'entraînement donc l'architecture du AG est la suivante : Global average pooling -> MLP -> ReLU ->FC -> Sigmoid.

2. Architecture et résultat

stage	output	OSNet
conv1	128×64, 64 64×32, 64	7×7 conv, stride 2 3×3 max pool, stride 2
conv2	64×32, 256	bottleneck × 2
transition	64×32, 256 32×16, 256	1×1 conv 2×2 average pool, stride 2
conv3	32×16, 384	bottleneck × 2
transition	32×16, 384 16×8, 384	1×1 conv 2×2 average pool, stride 2
conv4	16×8, 512	bottleneck × 2
conv5	16×8, 512	1×1 conv
gap	1×1, 512	global average pool
fc	1×1, 512	fc
# params		2.2M
Mult-Adds		978.9M

Table 1: Architecture of OSNet with input image size 256 × 128.

Method	Publication	Backbone	Market1501		CUHK03		Duke		MSMT17	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
ShuffleNet†‡ [72]	CVPR'18	ShuffleNet	84.8	65.0	38.4	37.2	71.6	49.9	41.5	19.9
MobileNetV2†‡ [38]	CVPR'18	MobileNetV2	87.0	69.5	46.5	46.0	75.2	55.8	50.9	27.0
BraidNet† [57]	CVPR'18	BraidNet	83.7	69.5	-	-	76.4	59.5	-	-
HAN† [27]	CVPR'18	Inception	91.2	75.7	41.7	38.6	80.5	63.8	-	-
OSNet† (ours)	ICCV'19	OSNet	93.6	81.0	57.1	54.2	84.7	68.6	71.0	43.3
DaRe [58]	CVPR'18	DenseNet	89.0	76.0	63.3	59.0	80.2	64.5	-	-
PNGAN [35]	ECCV'18	ResNet	89.4	72.6	-	-	73.6	53.2	-	-
KPM [41]	CVPR'18	ResNet	90.1	75.3	-	-	80.3	63.2	-	-
MLFN [2]	CVPR'18	ResNeXt	90.0	74.3	52.8	47.8	81.0	62.8	-	-
FDGAN [10]	NeurIPS'18	ResNet	90.5	77.7	-	-	80.0	64.5	-	-
DuATM [42]	CVPR'18	DenseNet	91.4	76.6	-	-	81.8	64.6	-	-
Bilinear [46]	ECCV'18	Inception	91.7	79.6	-	-	84.4	69.3	-	-
G2G [39]	CVPR'18	ResNet	92.7	82.5	-	-	80.7	66.4	-	-
DeepCRF [3]	CVPR'18	ResNet	93.5	81.6	-	-	84.9	69.5	-	-
PCB [47]	ECCV'18	ResNet	93.8	81.6	63.7	57.5	83.3	69.2	68.2	40.4
SGGNN [40]	ECCV'18	ResNet	92.3	82.8	-	-	81.1	68.2	-	-
Mancs [54]	ECCV'18	ResNet	93.1	82.3	65.5	60.5	84.9	71.8	-	-
AANet [50]	CVPR'19	ResNet	93.9	83.4	-	-	87.7	74.3	-	-
CAMA [65]	CVPR'19	ResNet	94.7	84.5	66.6	64.2	85.8	72.9	-	-
IANet [15]	CVPR'19	ResNet	94.4	83.1	-	-	87.1	73.4	75.5	46.8
DGNet [77]	CVPR'19	ResNet	94.8	86.0	-	-	86.6	74.8	77.2	52.3
OSNet (ours)	ICCV'19	OSNet	94.8	84.9	72.3	67.8	88.6	73.5	78.7	52.9

Table 3. Results (%) on big ReID datasets. It is clear that OSNet achieves state-of-the-art performance on all datasets, surpassing most published methods by a clear margin. It is noteworthy that *OSNet has only 2.2 million parameters*, which are far less than the current best-performing ResNet-based methods. -: not available. †: model trained from scratch. ‡: reproduced by us. (Best and second best results in red and blue respectively)

On constate qu'Osnet a atteint l'état de l'art avec des résultats très prometteurs sur différents jeux de données de reidentification. Les performances de ce modèle ont été évaluées sur plusieurs métriques couramment utilisées, telles que la précision, le taux de rappel, et le Mean Average Precision (MAP), et il a démontré une grande efficacité dans la tâche de reidentification des personnes en multivue. Les performances exceptionnelles d'Osnet peuvent s'expliquer par son architecture innovante et sa capacité à extraire des caractéristiques uniques et distinctes pour chaque personne à partir des images.

VI. Résultats expérimentaux

J'ai testé le modèle Omniscale, disponible dans la bibliothèque torchreid, qui permet de facilement entraîner

des modèles de Re-ID à partir de zéro en choisissant l'architecture et l'ensemble de données souhaités. La bibliothèque offre également des fonctionnalités pour tester les modèles entraînés. Bien que cette bibliothèque soit très utile, elle manque de documentation et d'exemples d'utilisation. J'ai personnellement eu besoin de temps pour la comprendre, ce qui m'a poussé à préparer une notebook expliquant comment tester des modèles de Re-ID avec cette bibliothèque.
Des modèles pré-entraînés sont également disponibles sur leur site, mais pour obtenir de meilleurs résultats, il est recommandé d'effectuer un tuning des hyperparamètres. En somme, la bibliothèque torchreid est un outil très pratique pour l'entraînement et le test de modèles de Re-ID, mais elle pourrait bénéficier d'une documentation et d'exemples plus complets pour faciliter son utilisation.

J'ai testé le modèle Osnet sur le dataset Market_15_01 en ajoutant mes propres photos pour évaluer ses performances. J'ai obtenu une précision de 82,5 % et un taux de réussite de 97,9 % pour le Rank_5, ce qui signifie que la personne recherchée était présente dans les cinq premiers. Voici les résultats obtenus :

```
** Results **
mAP: 82.5%
CMC curve
Rank-1  : 94.2%
Rank-5  : 97.9%
Rank-10 : 98.7%
Rank-20 : 99.2%
```


L'image de gauche est l'image requête et les cinq images de droite sont les cinq meilleurs résultats. La bordure verte indique une identification correcte de la même personne que dans l'image requête, tandis que la bordure rouge indique une identification incorrecte.

VII. Conclusion

En conclusion, la réidentification des personnes en multivue est un domaine de recherche complexe et en constante évolution. Les systèmes de Re-ID doivent faire face à des défis tels que la détection de personnes, la correspondance entre les sujets dans les images, les changements d'illumination, la faible résolution, l'occlusion, les vêtements uniformes, la scalabilité, entre autres. Malgré ces défis, la Re-ID continue d'attirer l'attention de la communauté de recherche en raison de son potentiel à améliorer la sécurité publique et la surveillance. De nombreuses avancées ont été réalisées dans ce domaine, mais il reste encore beaucoup de travail à faire pour améliorer la précision et la fiabilité des systèmes de Re-ID en multivue.

VIII. Bibliographie

https://kaiyangzhou.github.io/deep-person-reid/
https://arxiv.org/pdf/2001.04193.pdf
https://arxiv.org/pdf/2207.14452.pdf
https://arxiv.org/pdf/1905.00953.pdf
https://arxiv.org/pdf/2110.04764.pdf
https://zheng-lab.cecs.anu.edu.au/Project/project_reid.html
https://openai.com/
https://www.youtube.com/watch?v=WrfZIw1-zvc&t=859s
https://github.com/KaiyangZhou/deep-person-reid
https://github.com/layumi/Person_reID_baseline_pytorch