**Rapport 1er Jalon**

**Introduction :**

La réidentification de personnes est un domaine en pleine expansion, dont l'objectif est de reconnaître un individu à travers différentes caméras ou à différents moments. Dans le cadre de ce rapport, nous nous penchons sur les méthodes de réidentification présentées dans le survey qui nous a été attribué [1]. Nous explorerons les principales techniques, leurs caractéristiques, leurs avantages et inconvénients, ainsi que les datasets couramment utilisés dans ce domaine.

**Pedestrian description :**

La ré-identification s’effectue en plusieurs étapes. Supposé que l’on possède un dataset d’images ou de vidéos à analyser, la première étape est d’identifier et de reconnaître tous les individus présents sur nos images données en entrée et de les isoler. Pour cela, de nombreuses méthodes de détection existent afin d’isoler les cibles à analyser et identifier, voici une description rapide des plus connues et utilisées :

*Weighted Color Histogram (WCH) :* On isole d’abord le passant dans une image réduite, puis on essaie au plus possible de retirer l’arrière-plan pour ne conserver que le piéton en utilisant des axes de symétrie sur les parties de son corps et en donnant plus de poids aux pixels proches de ces axes. Cela permet de déterminer ce qui semble être en arrière-plan et que l’on peut retirer de l’analyse grâce à des probabilités. Enfin, on range les pixels par couleur afin de tirer un histogramme des couleurs dominantes les plus affluentes sur un cas précis.

*Maximally Stable Color Regions (MSCR) :* On groupe les pixels de couleurs similaires dans des petits clusters stables, pour ensuite pouvoir décrire les différentes régions et zones de l’image par leur taille, localisation, centre, matrice de moments, et couleur moyenne

*Recurrent Highly Structured Patches (RHSP) :* On utilise l’entropie pour détecter les textures avec des bords importants et à forte occurrence.

**Distance Metric Learning :**

La deuxième étape de la ré-identification consiste à mesurer à quel point l’individu observé est proche d’un autre individu déjà analysé. En effet, une bonne mesure de distance pour les fonctions de coût est importante car les traits de visage de chaque passant sont trop difficiles à identifier et reconnaître dans la masse de variances et de facteurs différents d’un échantillon à l’autre. Les méthodes d’apprentissage métriques sont catégorisées comme de l’apprentissage supervisé, contre l’apprentissage non supervisé. L’idée générale de ces méthodes est de conserver les vecteurs d’une même classe ensemble, en éloignant ceux de classes différentes.

La méthode la plus connue est la méthode KISSME, même s’ils en existent de nombreuses autres, dont des dérivations de cette dernière. Plus en détails, la méthode KISSME est basée sur cette équation représentant la distance carrée entre 2 vecteurs et et où M est une matrice semi-définie positive :

Avec cette méthode, on décide de si une paire est similaire ou non sur des probabilités. La différence de la paire est employée, et on assume que l’espace de différence serait de distribution Gaussienne avec une moyenne à 0.

**Systèmes de deep-learning :**

En deep-learning, un CNN (*convolutional neural network*) est une classe de réseaux neuronaux profonds, plus couramment appliqué pour analyser l'imagerie visuelle [2][5][6]. Ce procédé utilise la convolution qui est une opération mathématique prenant en entrée deux matrices, celle de l’image et du filtre appliqué et qui retourne une matrice exprimant la matrice convoluée. Dans l’identification de personne basée sur l’image, deux modèles de CNN sont communément adoptés : le siamois et la classification. Le manque de données fourni dans la réidentification et sa capacité d’apprendre à partir de très peu de données a rendu les réseaux siamois plus populaires ces dernières années. [4]. Il est composé d’une paire d’images en entrée partitionnée qui traverse différentes piles de couches convolutives utilisant chacune des filtres (convolutive layer,maximum pooling , fc layer) avec différentes fenêtres de pixels [3]. La similarité des deux vecteurs est par la suite calculée. L’un des points forts de cette méthode est que très peu d’images par classe sont suffisantes pour l’apprentissage de celle-ci et leur reconnaissance à l’avenir [4]. Néanmoins, comme le réseau siamois implique des paires quadratiques à partir desquelles se base l’apprentissage, il est plus lent lorsque l’on utilise de grand dataset. De plus, il ne fait pas une utilisation complète des annotations de réidentification ne considérant que les étiquettes par paires (ou triplet). D’une autre part le modèle de classification/ identification, performe très bien sur des dataset volumineux et fait une utilisation complète des étiquettes de réidentification. Des alternatives sont trouvées utilisant différents choix architecturaux et des stratégies pour remédier aux limites de chacun des modèles.

**Datasets et évaluation :**

Il existe plusieurs datasets d'images populaires pour la re-identification par image comme VIPeR qui contient 632 identités avec 2 images par identité capturées par 2 caméras différentes dans des conditions d'éclairage variables. D'autres datasets tels qu'iLIDS, CUHK et Market-1501 proposent également différents scénarios (hall d’aéroport, université) mais leurs tailles, bien que plus importantes dans les datasets récents avec plus de 1000 identités et 10 000 bounding boxes, restent insuffisantes pour entraîner des modèles robustes.

Dans ces datasets, les bounding boxes sont générées à la main pour les anciens et plus récemment automatiquement à l'aide de détecteurs de piétons comme DPM (Deformable Part Model) ou ACF (Aggregate Channel Features) afin d’économiser beaucoup de temps et de ressources humaines, mais cela peut introduire des erreurs de détection affectant les performances de re-ID.

Concernant les données vidéo, des datasets comme ETH, 3DPeS, iLIDS-VID et PRID-2011 capturant jusqu'à plus de 700 identités avec des séquences provenant de 1 à 8 caméras ont été développés. Le récent MARS contient quant à lui 1261 identités et plus de 20 000 séquences, les bounding boxes sont générées avec DPM et GMMCP tracker. ETH est arrivé à saturation avec quasiment 100% de précision, les autres sont en constante progression au fil des anneés.

La courbe de caractéristiques cumulatives (CMC) est largement utilisée pour évaluer les algorithmes, représentant la probabilité qu'une identité en requête apparaisse dans les k premiers candidats. Toutefois, elle ne prend en compte que la première correspondance, ce qui empêche de retourner tous les matchs dans le cas où une identité est capturée par plusieurs caméras (Ex : Market-1501 avec environ 14.8 vrais matchs cross-camera par requête) [7]. La précision moyenne (mAP) est plus complète que CMC en évaluant la capacité à trouver toutes les correspondances pour une requête. Pour comparer les performances entre plusieurs datasets, c’est la précision au rang 1 qui est prise en compte, c’est à dire si le premier candidat de la liste correspond à l’identité en requête

De manière générale, les approches d'apprentissage profond comme les CNN+RNN ont obtenu de meilleures performances que les méthodes traditionnelles, excepté pour le dataset VIPeR en raison de sa petite taille et de l'invariance de ses images.

**Video-based person Re-Id:**

Dans l’identification de personne basée sur la vidéo, il y a aussi les mêmes modèles qu’avec les images, c’est-à-dire le modèle de classification, et le modèle siamois. Cependant, un des défauts du modèle siamois, surtout avec une fonction de déclenchement (gating function) qui permet de capturer des patternes discrets, est que l’on doit pairer chaque image avec le reste de la database. Cela devient vite très long, et si ces modèles sont simples, ils manquent d’efficacité.

C’est pourquoi un nouveau modèle fut proposé : triplet loss function. Ce modèle utilise une image “ancre” pour comparer 2 autres images, une image positive et une négative. La positive est l’image qui présente la même personne que sur l’image ancre, et la négative une autre personne. Le but est que le vecteur CNN de l’image positive soit proche de celui de l’ancre et que la négative soit éloignée. L’entrainement du modèle nécessite des exemples complexes pour que l’IA arrive à identifier les paternes discrets, si les exemples sont trop simples, l’IA n’apprendra pas et sera bloquée dans une fausse direction. Il faut donc trouver des triplés ou la négative est proche de l’ancre. [8]

Des systèmes tels que Hand-crafted Systems, Multi-shot matching schemes et Spatial-temporal descriptors sont employés pour améliorer la précision.

Hand-crafted Systems: Ces systèmes sont conçus pour identifier les individus en se basant sur des caractéristiques spécifiques et définies manuellement.

Multi-shot matching schemes: Ces schémas utilisent plusieurs images ou séquences d'un individu pour améliorer la précision de la ré-identification.

Spatial-temporal descriptors: Ces descripteurs capturent les informations spatiales et temporelles d'une séquence vidéo pour aider à la ré-identification.

**Références bibliographiques :**

Person Re-identification: Past, Present and Future [1]

[CNN for Deep Learning | Convolutional Neural Networks (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/) [2]

[1601.07255.pdf (arxiv.org)](https://arxiv.org/pdf/1601.07255.pdf) [3]

[A friendly introduction to Siamese Networks | by Sean Benhur | Towards Data Science](https://towardsdatascience.com/a-friendly-introduction-to-siamese-networks-85ab17522942) [4]

<https://www.geeksforgeeks.org/introduction-convolution-neural-network/> [5]

<https://www.freecodecamp.org/news/convolutional-neural-network-tutorial-for-beginners/> [6]

<https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Zheng_Scalable_Person_Re-Identification_ICCV_2015_paper.pdf> [7]

<https://www.v7labs.com/blog/triplet-loss> [8]