Pertinence des combinaisons traqueur-détecteur pour le suivi-par-détection

G. Marion[1,2] A. A. Mekonnen[1] F. Lerasle[1,2]

[1] CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France
[2] Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France

{marion, aamekonn, lerasle}@laas.fr

Résumé

Ces travaux étudient la pertinence des associations filtre-détecteur dans toute stratégie visuelle de suivi par détection. Ils privilégient notamment les méthodes de Monte Carlo et des traqueurs basés sur des banques de représentations couplés à des détecteurs visuels. Les évaluations quantifient l'influence des associations filtre, détecteur, mono vs. multi représentations sur les performances globales. Nous observons, par exemple, que celles-ci sont très liées au choix du détecteur ; une différence de 1% *en rappel entre détecteurs peut induire une baisse de* 10% *sur la précision du traqueur global.*

Mots Clef

Suivi par détection, suivi multi-personnes, détection de personnes

1 Introduction

La détection et le suivi de personnes sont des domaines de recherche très actifs avec de multiples applications en vidéo-surveillance, pour les systèmes de protection des piétons, pour les interfaces homme-machine, en robotique... etc. C'est ainsi que ce domaine a attiré l'attention de la communauté de vision par ordinateur comme démontré par ces publications récentes : [1, 2, 3]. Le suivi de personnes dans un scénario donné, le suivi multi-personnes, est un sous-problème du suivi multi-objet (MOT - *Multi-Object Tracking*). Le suivi multi-objet tente d'estimer l'état (position, identité, configuration...) des objets présents dans la scène au cours du temps à partir des observations visuelles. En raison des difficultés classiques du domaine (encombrement de la scène, dynamique des cibles, variation intra- et inter-classe, bruit de mesure, mouvement du capteur, fréquence de trame) il est de notoriété publique que le couplage du système de suivi avec un détecteur, au sein d'un paradigme nommé *suivi par détection* (ou *tracking-by-detection*) améliore le suivi de manière significative [1, 4]. Dans le contexte de cette étude, les approches de suivi par détection s'appuient sur le détecteur de personnes afin d'initialiser, mettre à jour, réinitialiser, guider (c'est à dire éviter la dérive du suivi) ou arrêter un traqueur. Dans la littérature, de nombreuses approches de suivi par détection ont été appliquées au suivi de personnes. En revanche, la démarche classique est de choisir un détecteur prédéfini et de le coupler directement avec le traqueur (*e.g.*, [1, 5]), sans effectuer d'évaluations comparatives préliminaires. Suite aux récentes avancées en matière de détection de personnes, dont les performances ou la vitesse de détection varient fortement en raison d'avancées fulgurantes dans les domaines de l'apprentissage machine et de l'exploration de données [3, 2], et à l'asymétrie des progrès de la recherche en détection et suivi il est nécessaire que cela change. La première étape devrait être l'influence qu'a le choix du détecteur sur les performances du suivi ainsi que celle des associations entre les différents détecteurs et stratégies de suivi. A notre connaissance, aucune étude de cet ordre n'existe à l'heure actuelle. Il y a en effet d'excellente études comparatives dans le domaine de la détection, *e.g.*,[3], ainsi qu'en suivi *e.g.*,[6], mais aucune qui ne se focalise sur l'interdépendance entre choix du détecteur et du traqueur ainsi qu'à son effet dans différents contextes applicatifs.

Nous nous intéressons à trois stratégies de suivi (de filtrage) différentes, en raison de leur prévalence dans la littérature ainsi que leur pertinence : un Filtre Particulaire Décentralisé (DPF), *e.g.*,[1], le *Tracker Hierarchy* [7], et le *Reversible Jump Markov Chain Monte Carlo Particle Filter* (RJMCMC) [4]. DPF et RJMCMC sont sélectionnés car il s'agit des méthodes de Monte-Carlo les plus fréquentes dans la littérature, tandis que le *Tracker Hierarchy* permet l'étude d'un traqueur déterministe, donc complémentaire. Ces traqueurs sont couplés à quatre détecteurs parmi les plus fréquemment utilisés : HOG-SVM [8], DPM [9], ACF [2], et LDCF [10] (dont une présentation détaillée sera faite en section 3). Ce choix est motivé par le fait qu'ACF, LDCF et DPM font actuellement partie des meilleurs détecteurs. HOG-SVM en revanche, bien qu'éloigné de l'état de l'art en ce qui concerne les performances, utilise des descripteurs qui font partie d'une manière ou d'une autre de toutes les approches de l'état de l'art actuel. De plus, celui-ci a historiquement été le détecteur de référence pour les études comparatives.

Cet article est organisé comme suit. La section 2 détaille le système utilisé. Les sections 3 et 4 présentent respectivement les détecteurs et traqueurs sélectionnés. La section 5 détaille les expériences réalisées et les résultats obtenus. La section 6 ouvrira sur une discussion des résultats et sur les conclusions subséquemment tirées.

Travaux connexes Le suivi par détection est usuel en suivi multi-cibles, tout particulièrement dans les domaines de la vidéo-surveillance et autres systèmes de suivi de personnes en raison d'amélioration récentes des détecteurs d'objets cibles [6]. Celui-ci fournit une fondation fiable sur laquelle concevoir des traqueurs multi-cibles résistants à la dérive, à la perte de cible, aux occultations et à la confusion d'identité. De plus, celui-ci constitue le paradigme dominant pour ce qui est du suivi, en particulier en suivi multi-personnes [1]. L'avantage principal d'une telle méthode est qu'elle permet aux traqueurs multi-cibles de s'appuyer sur le détecteur pour initialiser, corriger ou terminer une trajectoire.

Les approches de suivi par détection de la littérature peuvent être grossièrement regroupées en deux catégories : traitement par lot ou méthodes en ligne. Le traitement par lot effectue une optimisation globale sur tout le contenu d'une vidéo afin d'inférer les trajectoires des cibles à partir des détections, *e.g.*,[11]. Les approches en ligne, elles, peuvent mettre en œuvre soit un processus de Markov de premier ordre, employant alors une stratégie probabiliste récursive pour ensuite préciser la trajectoire des cibles trame par trame *e.g.*,[1], ou bien générer des fragments de trajectoires (appelés *tracklets*) en liant les détections une à une trame après trame, pour ensuite les associer de manière plus globale afin de construire des trajectoires plus longues *e.g.*, [12]. Comme on peut le deviner étant donnés leurs noms respectifs, le traitement par lot est utilisé dans des contextes hors-ligne tandis que les méthodes en ligne sont courantes dans des contextes de suivi en temps réel. Dans cette publication, nous nous focaliserons sur les traqueurs à processus Markoviens de premier ordre.

Le paradigme du suivi par détection n'est devenu populaire qu'après de grandes avancées dans le domaine de la détection de personnes [13, 11]. Jusqu'à très récemment, la grande majorité des travaux en suivi par détection se sont appuyés sur le détecteur de personnes HOG-SVM [1, 7]. Encore aujourd'hui, en vertu du grand nombre de détecteurs disponibles, et dont les performances varient grandement, seule une petite portion d'entre eux, plus précisément HOG-SVM et ACF [1, 14], a été considérée dans le cadre du suivi par détection. Pourtant, la littérature concernant la détection de personnes est assez dense (voir [3] pour un recensement plus exhaustif), ce qui justifie la nécessité d'évaluations comparatives montrant l'importance du choix du détecteur sur les performances du traqueur dans différents contextes.

Contributions Nos contributions peuvent être résumées ainsi : (1) Extension de deux traqueurs courants utilisant les méthodes de Monte Carlo au suivi multi-représentations ; (2) Poursuite de nos évaluations systématiques [15] d'approches de suivi par détection sur des combinaisons traqueur-détecteur, en les étendant à d'autres détecteurs et contextes applicatifs ; (3) Présentation des résultats pertinents, et discussions associées, mettant en exergue l'importance des choix du traqueur, du modèle d'apparence et du détecteur.

2 Synoptique général

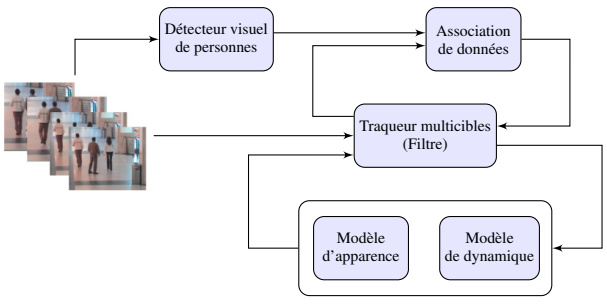

FIGURE 1 – Synoptique d'un système usuel de suivi par détection.

Le *framework* de suivi par détection sélectionné est détaillé en figure 1. Celui-ci met à profit le détecteur afin d'initialiser, terminer et mettre à jour les trajectoires. L'association de données fait correspondre trajectoires et détections, identifiant les détections comme soit une des cibles auquel cas celle-ci est utilisée afin de mettre à jour la trajectoire de cette cible, soit une nouvelle cible auquel cas elle pourra être utilisée pour initialiser une nouvelle trajectoire si celle-ci est associée à de futures détections. Le traqueur, ou filtre, se charge de la propagation des trajectoires à la trame en cours suivant un modèle de dynamique, et estime la position la plus probable en vertu du modèle d'apparence. Le rôle du filtre est de compenser le manque de fiabilité du détecteur, c'est-à-dire le grand nombre de faux positifs et de faux négatifs ainsi que le fait que sa réponse soit discrète plutôt, par-exemple, que probabiliste.

(a) HOG-SVM (b) DPM (c) ACF (d) LDCF

FIGURE 2 – Exemples de détections des quatre détecteurs.

La sortie du détecteur consiste en un ensemble de détections non étiquetées, ainsi que leur note de confiance, obtenues à partir d'une trame. La note de confiance du détecteur peut être seuillée afin de minimiser les fausses détections. La liste des détecteurs privilégiés est présentée en détail en section 3. Ces détections sont ensuite envoyées au module d'association de données qui fait correspondre les détections de la trame courante avec les trajectoires des trames précédentes s'il y en a en fonction de leurs positions, tailles et apparences grâce à un algorithme glouton détaillé dans [1]. Lorsqu'une détection est assignée à une trajectoire, la durée de vie de la trajectoire en question est augmentée, tandis que les trajectoires auxquelles aucune détection n'a été associée voient leur durée de vie diminuer. Lorsqu'une trajectoire n'est associée à aucune détection sur un certain nombre de trames, la trajectoire est supprimée. Les détections non-associées à des trajectoires sont utilisées pour créer des trajectoires "potentielles", qui deviendront eux-même des trajectoires actives lorsqu'elles seront associées avec un nombre suffisant de détections successives. Lorsque l'association de données est terminée, le suivi (filtrage) a lieu. La liste des traqueurs sélectionnés est détaillée en section 4. Il faut mentionner que toutes les tâches de suivi effectuées dans le cadre de cette étude sont faites dans le plan image, la sortie du traqueur décrivant alors la boîte englobante correspondant à la cible.

3 Détecteurs de personnes

Dans cette section nous présentons les différents détecteurs de personnes utilisés dans les évaluations. L'état de l'art dans le domaine des détecteurs de personnes regroupe différents détecteurs dont les performances en détection et temps de calcul ou le niveau d'abstraction du modèle varient grandement. Dans cette article, nous sélectionnons quatre détecteurs : HOG-SVM [8], DPM [9], ACF [2], et LDCF [10], pour nos évaluations en suivi par détection.

Histogram of Oriented Gradients (HOG-SVM) Ce détecteur, développé par Dalal et Triggs [8], est un des détecteurs les plus anciens et des plus classiques. Ce détecteur calcule les histogrammes locaux de l'orientation du gradient et utilise un séparateur à vaste marge (SVM) pour la classification. Le modèle appris est une abstraction holistique (comprenant le corps entier) entraînée sur le jeu de donnée INRIA [8]. Un exemple de détection est illustré par la figure 2a.

Deformable Parts Model (DPM) DPM [9] est un détecteur basé-parties qui agrège les indices de présence des différentes

parties afin de détecter une personne dans l'image. Le détecteur utilise conjointement un modèle global basé-parties et une version modifiée des descripteurs du HOG. L'apprentissage est effectué en entraînant un SVM sur un jeu de données partiellement labélisé de l'INRIA. Un exemple de détection est illustré par la figure 2b.

Aggregate Channel Features (ACF) ACF est un détecteur de personnes rapide basé sur la notion de canaux de descripteurs (*Channel Features*) qui a grandement amélioré en comparaison des détecteurs précédemment disponibles les performances en détection sur de nombreux jeux de données [2]. Il s'appuie sur des agrégats de descripteurs hétérogènes que l'on représente comme des canaux, un classifieur appris par *boosting*, et un modèle d'apparence holistique. ACF s'appuie sur dix canaux : valeur du gradient, HOG (sur six canaux), et canaux couleur LUV. Dans cette étude, ACF est entraîné sur le jeu de données INRIA.Un exemple de détection est illustré par la figure 2c.

Locally Decorrelated Channel Features (LDCF) LDCF [10] est un détecteur basé sur les canaux de descripteurs tout comme ACF. Cependant, plutôt que d'entraîner le classifieur directement sur les descripteurs il leur applique une étape de décorrélation pour chaque canal. Un exemple de détection est illustré par la figure 2d.

LDCF et ACF, comparativement aux autres détecteurs de cette étude, obtiennent de bien meilleures performances dans des environnements extérieurs, comme par-exemple sur des caméras montées sur un véhicule, alors que DPM est plus performant sur des jeux de données contenant des occurrences d'occultations partielles.

4 Suivi multi-cibles (MOT)

Nous nous intéressons au MOT dans le cadre du suivi multi-personnes grâce à des méthodes de suivi par détection comme vu en section 2. Deux types de filtres retiennent notre attention : ceux purement probabilistes, basés sur des méthodes de Monte-Carlo (DPF et RJMCMC), et une approche déterministe du type Tracker Hierarchy. DPF et RJMCMC sont sélectionnés car ils représentent deux configurations de filtres usuelles : centralisée et décentralisée. Nous leur associons tout d'abord d'un modèle d'apparence multi-représentations inspiré de celui du Tracker Hierarchy afin de les évaluer sur une base commune. Une fois que les trois filtres utilisent des modèles d'apparence similaires nous les évaluons sur des jeux de données publics (Section 5) en les combinant tour-à-tour avec chacun des quatre détecteurs.

4.1 Decentralized Particle Filter (DPF)

En MOT décentralisé, chaque filtre instancié a son propre vecteur d'état indépendant de ceux des autres filtres. Pour ce type de filtres, nous choisissons d'évaluer notre variante du filtre particulaire (PF) classique, un choix fréquent dans la littérature, *e.g.*, [1]. Dans ce cadre, chaque cible se voit attribuer une instance unique d'un filtre particulaire pour le suivi. Ce filtre est inspiré par ICONDENSATION [16], une méthode de Monte Carlo séquencielle permettant d'approximer la probabilité à posteriori de l'état x_t de la cible en s'appuyant sur les mesures $Z_{1:t}$ faites jusqu'à l'instant t, et étendue au MOT. Il se base pour cela sur un ensemble de N particules, *i.e.*, $p(x_t|Z_{1:t}) \approx \{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N$. Le suivi se fait séquenciellement grâce à l'échantillonnage pondéré selon lequel les particules à l'instant $t-1$ sont propagées à l'instant t en suivant une densité $q(.)$, et leurs poids mis à jour selon l'équation 1.

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(z_t|x_t^{(i)})p(x_t^{(i)}|p(x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{t-1}^{(i)}, z_t)} \quad (1)$$

Où $p(x_t^{(i)}|x_{t-1}^{(i)})$ est le modèle de dynamique de la cible, $p(z_t|x_t^{(i)})$ le terme de vraisemblance, et $q(x_t^{(i)}|x_{t-1}^{(i)}, z_t)$ la fonction d'échantillonage évaluée à $t-1$. La fonction d'échantillonage de l'équation 2 permet de combiner ce filtre avec les détections dans le cadre du suivi par détection : une partie des particules seront échantillonnées à partir des détections, $\pi(x_t^{(i)}|z_t)$, d'autres à partir du modèle de dynamique, et certaines à partir de l'a-priori $p_0(x_t^{(i)})$ selon les rapports respectifs α, β, et γ dont la somme doit être unitaire.

$$q(x_t^{(i)}|x_{t-1}^{(i)}, z_t) = \beta p(x_t^{(i)}|x_{t-1}^{(i)}) + \alpha\pi(x_t^{(i)}|z_t) + \gamma p_0(x_t^{(i)}) \quad (2)$$

L'état x_t d'une cible à l'instant t est représenté par le vecteur $[\mathsf{x}_t, \mathsf{y}_t, \mathsf{s}_t]^T$, où x et y représentent la position dans le plan image, et s représente l'échelle de la boîte englobante encadrant la cible à l'instant t. $p(z_t|x_t^{(i)})$ correspond à la mesure de vraisemblance obtenue en accord avec le modèle d'apparence. Le filtre particulaire classique met à jour l'apparence de la cible via un simple histogramme de couleur. Nous proposons ici une variante multi-représentations, appelée DPF-MT pour *Multi-Template*, prenant en compte plusieurs histogrammes. Un histogramme est calculé pour chacune des représentations enregistrées.

DPF (Mono-template) Celui-ci possède un simple histogramme de couleur 2D dans l'espace colorimétrique HS. La cible j à l'instant t est alors représentée par $\mathcal{H}_t^j$. L'histogramme $\mathcal{H}_t^{x_t}$ peut être dérivé de x_t et de l'image source : la mesure de vraisemblance peut alors être calculée en utilisant la distance de Bhattacharyya $\mathcal{B}(,)$ (voir Eq. 3).

$$p(z_t|x_t^{(i)})|_j \propto exp\left(-\lambda\mathcal{B}\left(\mathcal{H}_t^j, \mathcal{H}_t^{x_t}\right)\right) \quad (3)$$

$$\mathcal{H}_t^j = \alpha_h\mathcal{H}_t^{\bar{x}_t} + (1-\alpha_h)\mathcal{H}_{t-1}^i \quad (4)$$

H_t^j est mis à jour dynamiquement selon l'équation 4 avec $\mathcal{H}_t^{\bar{x}_t}$, l'histogramme calculé à la position $\bar{x}_t$ obtenue par minimisation de l'erreur quaratique moyenne.

DPF-MT (Multi-template) La variante multi-représentations met en jeu les histogrammes d'un ensemble de motifs de manière similaire à [7]. Dans ce cas, l'apparence de la cible est représentée par le jeu d'histogrammes $\left\{\mathcal{H}_{\mathsf{c}_k,k}^j\right\}_{k=1}^{N_T}$ où c_k représente les deux canaux du $k^{ième}$ histogramme. Ce changement effectué, la mesure de vraisemblance est donnée par l'équation 5.

$$p(z_t|x_t^{(i)})|_j \propto exp\left(-\lambda\sum_{k=1}^{N_T}\mathcal{B}\left(\mathcal{H}_{\mathsf{c}_k,k}^j, \mathcal{H}_{\mathsf{c}_k,t}^{x_t}\right)\right) \quad (5)$$

La mise à jour des représentations est gérée similairement à [7] : chaque détection associée à la cible ajoute une nouvelle représentation au filtre. Les deux canaux sont sélectionnés parmi les espaces colorimétriques RGB, HSV, Lab selon leur pouvoir discriminant entre la cible et l'arrière plan. Celui-ci est calculé comme la différence des scores en rétroprojection entre les deux régions. Lorsqu'une nouvelle représentation est ajoutée, les modèles les moins discriminants sont supprimés afin de ne conserver qu'au plus N_T représentation.

Dans la section 5 les deux filtres présentés ici sont évalués après couplage avec chacun des quatre détecteurs présentés en section 3.

4.2 Reversible Jump Markov Chain Monte Carlo - Particle Filter (RJMCMC)

RJMCMC [4] est un filtre centralisé encapsulant l'état de toutes les cibles dans un seul vecteur d'état. RJMCMC repose sur une chaîne de Markov sur les configurations d'état qui, après convergence, approxime la distribution a posteriori. L'échantillonnage d'importance est remplacée par un échantillonnage Metropolis Hastings (MH) sur la chaîne.
De même que DPF, RJMCMC modélise la densité a posteriori X_{t-1} à partir de toutes les mesures jusqu'à l'instant $t-1$ grâce à un ensemble de M particules. Cependant les particules ne sont pas pondérées, *i.e.*, $p(X_{t-1}|Z_{1:t-1}) \approx \{X_{t-1}^{(i)}\}_{i=1}^{M}$. De plus le vecteur d'état d'une cible j dans la particule i à l'instant t est le vecteur $x_{j,t}^i = [Id_{j,t}^i, \mathsf{x}_{j,t}^i, \mathsf{y}_{j,t}^i, \mathsf{s}_{j,t}^i]^T$, où $\mathsf{x}, \mathsf{y}, \mathsf{s}$ dénote la position et l'échelle de la cible dans le plan image, et Id dénote l'identifiant de la cible. Subséquemment, la $i^{ième}$ particule à l'instant t est représentée comme $X_t^i = \{I_t^i, x_{(j,t)}^i\}, j \in \{1, ..., I_t^n\}$, où I_t^i est le nombre de cibles modélisées par la $i^{ième}$ particule. RJMCMC modélise un ensemble de transitions m afin de changer la dimension du vecteur d'état *i.e.*, ajouter une nouvelle cible, cesser le suivi d'une cible, ou le prédire. Chaque transition m est associée à une fonction de proposition $Q_m(.)$, et doit avoir une transition inverse $m*$ permettant la réversibilité afin de faire converger la chaîne vers la distribution stationnaire désirée[4]. Lors de l'estimation itérative, à la $i^{ième}$ itération RJMCMC échantillonne une transition et propose une nouvelle particule x^* selon $Q_m(.)$. Le taux d'acceptation α_a (équation . 6) est alors calculé selon $Q_m(.)$ et $Q_{m^*}(.)$. x^* a une probabilité α_a d'être acceptée, ou bien est rejetée. Les particules utilisées pour la phase de *burn-in* (M_b) et de *thin-out* (M_{th}) sont retirées, laissant alors M échantillons non pondérés pour représenter la densité a posteriori.

$$\alpha_a = \min\left(1, \frac{p(X^*|Z_{1:t})Q_{m^*}(X_t^{(i-1)}; X^*)q_{m^*}\Psi(X^*)}{p(X_t^{(i-1)}|Z_{1:t})Q_m(X^*; X_t^{(i-1)})q_m\Psi(X_t^{(i-1)})}\right) \quad (6)$$

$\Psi(.)$ (Eq. 6) est le modèle d'interactions. Notre implémentation est basée sur [4] avec comme transitions $m = \{add, delete, stay, leave, update, swap\}$, et un modèle d'interactions basé sur un champ aléatoire de Markov $\Psi(.)$. Les différentes distributions spécifiques à chaque transition $Q_m(.)$ sont définies suivant [4]. A l'instar du DPF, nous proposons deux variantes du RJMCMC, fonctions du modèle d'apparence sous-jacent : une version mono-représentation RJMCMC, et une version multi-représentations RJMCMC-MT.

RJMCMC (mono-représentation) Ce filtre représente la cible par un histograme 2D, tout comme DPF. La seule différence est que $p(z_t|X_t^{(i)})$ est calculé pour les $i^{ièmes}$ particules puisque $X_t^{(i)}$ modélise les états de toutes les cibles. Soit le sous-ensemble $\hat{X}$ des cibles mises à jour ou échangées (à l'exception des cibles retirées ou ajoutées dont la vraisemblance vaut un). Alors, la vraisemblance de $X_t^{(i)}$ est évaluée via Eq. 7.

$$p(z_t|X_t^{(i)}) \propto exp\left(-\lambda \sum_{x \in \hat{X}} \mathcal{I}(j,x)\mathcal{B}\left(\mathcal{H}_t^j, \mathcal{H}_t^x\right)\right) \quad (7)$$

La fonction caractéristique $\mathcal{I}(j,x) = 1$ si $j = id(x)$, et 0 sinon garantit que l'apparence d'une cible est associée à la cible correspondante du vecteur d'état.

RJMCMC-MT (multi-représentations) De même que DPF-MT, la variante multi-représentations possède un jeu de N_T histogrammes de couleur. Avec $\hat{X}$ le sous-ensemble des cibles mises à jour ou échangées, le terme de vraisemblance est évalué selon Eq. 8.

$$p(z_t|X_t^{(i)}) \propto exp\left(-\lambda \sum_{x \in \hat{X}} \mathcal{I}(j,x) \sum_{k=1}^{N_T} \mathcal{B}\left(\mathcal{H}_{\mathsf{c}_k,k}^j, \mathcal{H}_{\mathsf{c}_k,t}^x\right)\right) \quad (8)$$

RJMCMC est couplé aux différents détecteurs présentés puis est évalué. Le couple filtre-détecteur est dénoté en apposant le nom du détecteur à RJMCMC *e.g.*, RJMCMC-ACF. Comme pour DPF l'association détection-trajectoire est gérée par un algorithme glouton. L'état est inféré par un estimateur type MMSE à partir du sous ensemble des particules encodant le nombre de cibles le plus représenté dans la population.

4.3 Tracker Hierarchy (Hierarchy)

Ce filtre multi-cibles est une autre approche décentralisée de suivi par détection qui assigne un filtre à chaque cible. Il met en jeu un ensemble de modèles de la cible pour modéliser son apparence, et forme une hiérarchie de filtres novices et experts pour un suivi multi-cibles efficace. Tracker Hierarchy alterne l'estimation de mode par *mean-shift* (prenant en compte l'apparence de la cible), et un filtre de Kalman (prenant en compte la dynamique de la cible) [7]. Nous l'évaluons après association avec différents détecteurs car, contrairement aux autres filtres, il adopte une approche déterministe et stochastique. Voir [7] pour plus de détails.

5 Evaluations et Résultats

Cette section présente les différentes évaluations, ainsi que les métriques utilisées, jeux de données sélectionnés, les détails d'implémentation et les résultats obtenus.

5.1 Métriques

Pour quantifier les performances des différents systèmes de suivi, nous adoptons le jeu de métriques classiques CLEAR-MOT [17]. CLEAR-MOT est utilisé principalement pour deux de ses composantes : le *Multi-Object Tracking Accuracy* (MOTA) et le *Multi-Object Tracking Performance* (MOTP). Une boîte englobante R_T estimée par le traqueur est considérée comme correcte si son taux de chevauchement $sc = \frac{R_T \cap G_T}{R_T \cup G_T}$, avec G_T la réalité terrain, est au-dessus d'un seuil τ, généralement choisi à 0.5.

5.2 Jeux de données

Vidéo	Caméra	Résolution	ips	#Trames	#Ids	Environnement
CAVIAR-EnterExit [a]	statique	384×288	25	383	4	intérieur (corridor)
CAVIAR-OneShop[a]	statique	384×288	25	1377	6	intérieur (corridor)
PETS-S2L1 [b]	statique	768×576	7	795	19	extérieur
TUD-Crossing [c]	statique	640×480	25	200	13	extérieur (route)
ETH-Bahnhoff [d]	mobile	640×480	14	1000	224	extérieur (trottoir)
ETH-Sunnyday[d]	mobile	640×480	14	354	30	extérieur (trottoir)
ETH-Jelmoli[d]	mobile	640×480	14	440	74	extérieur (trottoir)

TABLE 1 – Vidéos sélectionnées.

a. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
b. http://www.cvg.reading.ac.uk/PETS2009/a.html
c. http://datasets.d2.mpi-inf.mpg.de/andriluka08cvpr/
d. https://data.vision.ee.ethz.ch/cvl/aess/dataset/

Pour nos évaluations, nous utilisons sept jeux de données publics détaillés dans le tableau 1. Ces vidéos sont sélectionnées afin de recouvrir de nombreux scénarios, caractéristiques des cibles et configurations de la caméra, comme par-exemple : caméra fixe vs. mobile, résolutions variées, scénarios en intérieur vs. extérieur, arrière-plans encombrés ou dégagés, occultations, interactions entre cibles.

5.3 Détails d'implémentation [1]

L'évaluation des combinaisons détecteur-traqueur comprend un réglage attentionné des différents paramètres libres des filtres et détecteurs (seuils des détecteurs, nombre de particules, ...etc). A cet effet, nous constituons un jeu de données destiné au réglage composé des premières 20% images des vidéos PETS-S2L1, CAVIAR-OneShop, et ETH-Bahnof ($\sim$600 trames). Tous les paramètres, à l'exception de ceux tirés de travaux précédemment pubiés, sont réglés à partir de cette vidéo. Ce réglage est global. Les résultats expérimentaux finaux sont ensuite dérivés du reste des jeux de données (tous sauf celui de réglage). Nous préférons cette approche à un réglage spécifique à chaqe jeu de données afin de mieux refléter les performances globales des traqueurs indépendamment du domaine ou du contexte.

Les détecteurs utilisés pour les évaluations sont basés sur des implémentations libres publiquement disponibles ! LDCF et ACF, basés sur la librairie Matlab Dollar's Toolbox [3] ; DPM, basé sur son implémentation Matlab [9] ; et le détecteur HOG-SVM d'OpenCV (`http://docs.opencv.org/`) comprenant quelques modifications afin de fournir des cartes de scores au lieu d'une sortie binaire. Les deux traqueurs basés filtres à particules, DPF et RJMCMC, proviennent de nos implémentations (en `C++`). Le Tracker Hierarchy provient de l'implémentation originale en `C++` de [7], les paramètres autres que la fréquence d'images étant conservés.

5.4 Evaluations et résultats

Tous les résultats expérimentaux sont présentés ci après. Chaque combinaison détecteur-trauqueur est lancée dix fois sur chacun des sept vidéos sélectionnées. Ceci permet de prendre en compte la stochasticité des filtres particulaires et d'obtenir des résultats plus significatifs. Les métriques CLEAR-MOT MOTA et MOTP sont utilisées pour la présentation des résultats. Ici, plusieurs tableaux récapitulatifs présentant différents points de vue sur les résultats sont présentés [2]. Ceux-ci sont classés afin de rendre des comparaisons spécifiques plus faciles. Les catégories retenues tentent de répondre aux questions suivantes : (1) Les modèles multi-représentations sont-ils meilleurs (en performances et répétabilité) que ceux mono-représentation ? (2) Quelle combinaison détecteur-traqueur est la plus performante ? (3) Quel filtre, indépendamment du détecteur, obtient les meilleures performances ? (4) Quel détecteur, indépendamment du filtre, obtient les meilleures performances ?

(1) Approches Mono- et Multi-représentations . Afin de comparer les avantages à l'utilisation d'un modèle d'apparence multi-représentations sur un modèle mono-représentation, le tableau 2 présente un résumé des résultats obtenus avec les filtres particulaires. Les résultats sont moyennés sur tous les jeux de données et les types de détecteurs. Par exemple, ceci implique `10x4x7=280` exécutions de DPF-MT.

	DPF-MT	DPF	RJMCMC-MT	RJMCMC	Multi Représentations	Mono Représentation
MOTP↑	**.728**	**.728**	.645	.661	.687	<u>.694</u>
MOTA↑	**.415**	.409	.292	.244	<u>.354</u>	.327

TABLE 2 – Résultats mono- et multi-représentations pour les filtres particulaires, moyennés sur l'ensemble des jeux de données et des détecteurs.

Les résultats mettent en évidence que DPF-MT obtient en moyenne le meilleur MOTA avec 41.5%, le taux de vrais positifs étant également le meilleur dans ce cas. RJMCMC-MT obtient un faible score en vrais positifs. En moyenne, les modèles multi-représentations obtiennent en moyenne une amélioration de 3% en MOTA par-rapport aux modèles mono-représentation, avec de meilleurs taux de vrais positifs et de faux positifs. En termes de précision en revanche, les modèles mono-représentations obtiennent de légèrement meilleures performances, probablement dues à l'adaptation plus rapide du modèle à l'apparence de la cible.

(2) Evaluations générales du suivi par détection. Afin d'évaluer les performances globales en suivi par détection, le tableau 3 présente les résultats en MOTA et MOTP moyennés des approches multi-représentations pour chaque vidéo. Pour cinq des sept jeux de données, DPF-MT combiné au LDCF (DPF-LDCF-MT) obtient les meilleurs résultats, DPF-ACF-MT obtenant la seconde dans quatre des sept vidéos. D'autre part, Hierarchy-LDCF et RJMCMC-DPM-MT obtiennent respectivement les meilleurs MOTA sur PETS-S2L1 et CAVIAR-OneShop. Les résultats obtenu avec Hierarchy indiquent qu'il excelle sur PETS-S2L1 et CAVIAR-EnterExit mais obtient

1. Des détails sur les réglages des paramètres et leurs valeurs sont disponibles à l'adresse : http ://homepages.laas.fr/aamekonn/rfia_2016/

2. Les résultats bruts des évaluations sont disponibles à l'adresse `http://homepages.laas.fr/aamekonn/rfia_2016/`

de piètres performances sur les autres vidéos, tout spécialement lorsque le détecteur est HOG-SVM. Ceci tend à montrer que Hierarchy est performant sur les caméras statiques ambiantes.

Dataset	LDCF			ACF			DPM			HOG-SVM		
	I	II	III	I	II	III	I	II	III	I	II	III
PETS-S2L1	.770	.513	**.807**	.756	.483	.798	.637	.420	.728	.594	.164	.672
CAVIAR-OneShop	.371	.373	-.125	.343	.377	-.125	.351	**.422**	.300	.106	.171	-.347
CAVIAR-EnterExit	**.666**	.621	.647	.619	.544	.438	.518	.520	.510	.424	.423	.293
ETH-Bahnhof	**.375**	.259	.022	.281	.237	-.035	.239	.201	.109	.067	-.004	-.382
ETH-Jelmoli	**.364**	.296	.165	.357	.307	.124	.178	.155	.141	-.100	-.394	-.685
ETH-Sunnyday	**.621**	.423	.263	.560	.370	-.114	.469	.330	.402	.196	-.139	-.525
TUD-Crossing	**.650**	.484	.211	.568	.460	.207	.383	.363	.296	.251	-.193	-.130
Average	**.545**	.424	.337	.498	.397	.185	.396	.344	.355	.220	.004	-.158

Dataset	LDCF			ACF			DPM			HOG-SVM		
	I	II	III	I	II	III	I	II	III	I	II	III
PETS-S2L1	.723	.677	.761	.723	.674	.757	.723	.656	**.770**	.731	.586	.754
CAVIAR-OneShop	.717	.654	.735	.718	.647	.735	.743	.637	.730	.717	.570	**.758**
CAVIAR-EnterExit	.752	.729	.797	.753	.737	.803	.757	.695	**.842**	.751	.699	.795
ETH-Bahnhof	.725	.655	**.769**	.720	.640	.755	.720	.622	.766	.684	.557	.684
ETH-Jelmoli	.739	.640	.736	.734	.642	.740	.722	.602	**.742**	.674	.558	.673
ETH-Sunnyday	.773	.709	.770	.767	.707	.721	.757	.651	**.797**	.714	.525	.663
TUD-Crossing	.724	.697	.797	.733	.685	**.816**	.703	.647	.773	.697	.569	.690
Average	.736	.680	.766	.735	.676	.761	.732	.644	**.774**	.710	.581	.717

TABLE 3 – MOTA en suivi par détection [en haut] et MOTP [en bas]. I - DPF-MT, II - RJMCMC-MT, and III - Hierarchy. Le **meilleur** et le **second meilleur** résultats sont mis en valeur.

En moyenne, pour DPF-MT et RJMCMC-MT, le MOTA décroit dans l'ordre LDCF, ACF, DPM, HOG-SVM. Pour Hierarchy en revanche l'ordre est DPM, LDCF, ACF, HOG-SVM. La variation en MOTA entre le meilleur résultat et le moins bon est de 32.5%, 42%, et 51.3% respectivement pour DPF-MT, RJMCMC-MT, et Hierarchy. Hierarchy est clairement plus sensible au choix du détecteur. Le tableau 3 montre que Hierarchy obtient les meilleurs MOTP sur tous les jeux de données.

Vidéo	DPF-MT		RJMCMC-MT		Hierarchy	
	MOTP↑	MOTA↑	MOTP↑	MOTA↑	MOTP↑	MOTA↑
PETS-S2L1	.725	.689	.648	.395	**.761**	**.751**
CAVIAR-OneShop	.724	.293	.627	**.336**	**.739**	-.074
CAVIAR-EnterExit	.753	**.557**	.715	.527	**.809**	.472
ETH-Bahnhof	.712	**.241**	.619	.173	**.744**	-.072
ETH-Jelmoli	.717	**.200**	.610	.091	**.723**	-.064
ETH-Sunnyday	**.753**	**.462**	.648	.246	.738	.007
TUD-Crossing	.714	**.463**	.649	.279	**.769**	.146
Overall Average	.728	**.415**	.645	.292	**.754**	.179

TABLE 4 – Comparaison des performances en suivi sur chaque vidéo moyennées sur l'ensemble des détecteurs. Le **meilleur** résultat sur chaque vidéo est mis en valeur

(3) Évaluation des filtres. Le tableau 4 présente les résultats en MOTA et MOTP de chaque famille de filtres (en incluant les résultats en mono-représentation pour l'exhaustivité), moyennés sur l'ensemble des détecteurs, pour chaque vidéo. Les résultats montrent que DPF-MT obtient un meilleur MOTA sur cinq des vidéos. Pour les deux autres, Hierarchy et RJMCMC-MT obtiennent une marge d'avance de 6% et 4% respectivement sur les autres. En précision, Hierarchy obtient de meilleurs résultats sur six des jeux de données. Il devient clair que DPF-MT obtient de meilleures performances en filtrage indépendamment du couplage avec un détecteur, tandis que Hierarchy est bien meilleur pour la localisation de la cible suivie. Cela se manifeste également dans les résultats moyennés sur tous les jeux de données.

(4) Évaluation des détecteurs. Les résultats présentés dans le tableau 5 montrent les MOTA obtenus pour chacun des détecteurs moyennés sur l'ensemble des filtres – DPF-MT, RJMCMC-MT et Hierarchy. Pour faciliter la comaraison, la précision et le rappel (R/P) de chaque détecteur est également indiqué pour chaque vidéo. En termes de MOTA, pour six des sept jeux de données LDCF permet d'obtenir les meilleurs résultats. Pour le dernier, le meilleur résultat est obtenu avec DPM. Sur l'ensemble des vidéos, HOG-SVM obtient les moins bonnes performances. Étant donné que LDCF possède le meilleur rappel dans quatre des vidéos, l'amélioration en MOTA est atendue. Pour ETH-Jelmoli et ETH-Sunnyday, DPM possède le meilleur rappel, mais cela n'amène pas à obtenir le meilleur MOTA. Evidemment, comme la précision est inversement proportionnelle au taux de faux positifs, celle-ci joue un rôle important dans l'amélioration du MOTA. Une autre observation cruciale est la répétabilité des résultats obtenus avec DPM : pour toutes les vidéos, les résultats obtenus avec DPM ont le plus faible écart-type ce qui donne une indication de la répétabilité des résultats. Les résultats moyennés sur l'ensemble des vidéos montrent que les performances décroissent dans l'ordre suivant : LDCF, ACF, DPM, HOG-SVM. Malgré qu'il n'y ait pas de différence majeure en rappel et en précision entre LDCF et ACF, les MOTA correspondants diffèrent en moyenne de 7.5%. DPM obtient un MOTA meilleur de 0.5% comparé à l'ACF mais 7% inférieur à celui du LDCF. HOG-SVM quand à lui conduit au MOTA moyen le plus faible, 2.2%.

Dataset	LDCF		ACF		DPM		HOG-SVM	
	R/P	MOTA	R/P	MOTA	R/P	MOTA	R/P	MOTA
PETS-S2L1	**.92**/**.95**	**.696**	**.92**/.94	.679	.85/**.95**	.595	.81/.89	.477
CAVIAR-OneShop	.45/.93	.330	.43/**.95**	.192	**.51**/.88	**.358**	.37/.76	-.023
CAVIAR-EnterExit	**.71**/**.97**	**.644**	.69/.94	.534	.69/.89	.516	.62/.84	.380
ETH-Bahnhof	**.70**/**.83**	**.219**	.63/.78	.161	.57/.73	.183	.52/.47	-.107
ETH-Jelmoli	.54/**.90**	**.275**	.52/.87	.263	**.56**/.79	.158	.43/.52	-.393
ETH-Sunnyday	.70/**.91**	**.436**	.62/**.91**	.272	**.76**/.85	.400	.67/.60	-.156
TUD-Crossing	**.75**/**.93**	**.448**	.74/**.93**	.412	.63/.87	.347	.53/.68	-.024
Average MOTA	–	**.435**	–	.360	–	.365	–	.022

TABLE 5 – Résultats du suivi moyennés sur l'ensemble des filtres multi-représentations (DPF-MT, RJMCMC-MT, Hierarchy). R/P dénote la précision et le rappel du détecteur sur la vidéo donnée. Les **meilleurs** rappel, précision et MOTA sont indiqués.

6 Discussions et conclusions

Les résultats présentés ici permettent une bonne compréhension des performances de chaque combinaison détécteur-traqueur sur les différents jeux de données (donc contextes). De toute évidence, les résultats de chaque approche de suivi par détection est très influencée par le choix du détecteur. Commençons par évaluer les différences de performances entre les modèles mono- et multi-représentations pour DPF et RJMCMC à l'aide du tableau 2. En moyenne, les approches multi-représentations conduisent à une amélioration en MOTA mais pas en MOTP. Ceci peut être dû au fait que les approches mono-représentations adaptent leur modèle très rapidement en fonction de la dernière position estimée. Cela permet une localisation très précise de la cible mais ouvre la voie à la dérive du traqueur et *in fine* à la perte de la cible, d'où un MOTA plus faible.

Intéressons-nous maintenant au filtre qui obtient les meilleurs performances indépendamment du détecteur utilisé. Le tableau 4 des filtres basés multi-représentations montre que DPF-MT prend la tête en termes de MOTA. Il s'agit de loin du meilleur filtre, indépendamment du contexte. Celui-ci est suivi par RJMCMC-MT puis Hierarchy. En revanche, en termes de précision du suivi, Hierarchy est bien meilleur : l'estimation par *mean shift* permet une meilleure localisation de la cible. De même, le tableau 5 montre que LDCF est le meilleur détecteur. Il est suivi dans l'ordre par DPM, ACF puis HOG-SVM. Alors même qu'ACF suit directement LDCF en termes de détection dans l'état de l'art [10], DPM résulte en de légèrement meilleurs résultats. Ainsi, un meilleur détecteur en termes de détection n'implique pas nécessairement de meilleures performances en suivi. HOG-SVM, lui, apporte en moyenne un MOTA plus faible que ceux de tous les autres. Étant donné que l'état de l'art en suivi par détection repose sur des détecteurs basés sur HOG-SVM [7, 1] il est possible de grandement améliorer les résultats en utilisant LDCF (ou même ACF et DPM).

En résumé, de manière générale il est avantageux d'utiliser DPF-LDCF-MT en vertu de sa supériorité en termes de MOTA combinée à sa précision correcte. Si la caméra est fixe avec fort effet perspectif, donc si les déplacements sont faibles dans le plan image, Hierarchy-LDCF est conseillé. Lorsque la précision du suivi est importante et que le point de vue n'est pas fronto-parallèle, Hierarchy-DPM peut être utilisé. Les variantes de RJMCMC-MT ne doivent être privilégiées qu'afin de maintenir des performances similaires dans des contextes divers. De plus, il est crucial d'évaluer les couplages avec chaque nouveau choix de détecteur, car le rappel et la précision n'indiquent pas nécessairement le meilleur détecteur en termes de suivi, une différence de 1% en rappel pouvant amener à presque 10% de différence en MOTA. Ainsi, le choix du détecteur doit être plus mûrement réfléchi que le choix du filtre. Plus généralement, nous constatons que le choix du filtre et du détecteur n'améliore pas toutes les métriques dans tous les contextes, ce qui serait trop simple.

Références

[1] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. Van Gool, Online multiperson tracking-by-detection from a single, uncalibrated camera, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1820–1833.

[2] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.

[3] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection : An evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.

[4] Z. Khan, T. Balch, T. Dellaert, Mcmc-based particle filtering for tracking a variable number of interacting targets, IEEE Trans. Pattern Anal. Mach. Intell. 27 (11) (2005) 1805–1918.

[5] Y. Li, H. Ai, T. Yamashita, S. Lao, M. Kawade, Tracking in low frame rate video : A cascade particle filter with discriminative observers of different life spans, IEEE Trans. Pattern Anal. Mach. Intell. 30 (10) (2008) 1728–1740.

[6] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking : an experimental survey, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1442–1468.

[7] J. Zhang, L. L. Presti, S. Sclaroff, Online multi-person tracking by tracker hierarchy, in : IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'12), Beijing, China, 2012.

[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in : IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05), CA, USA, 2005.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[10] W. Nam, P. Dollár, J. H. Han, Local decorrelation for improved pedestrian detection, in : Advances in Neural Information Processing Systems (NIPS'14), 2014, pp. 424–432.

[11] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, in : Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[12] B. Wang, G. Wang, K. L. Chan, L. Wang, Tracklet association with online target-specific metric learning, in : IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14), 2014, pp. 1234–1241.

[13] W. Abd-Almageed, M. Hussein, M. Abdelkader, Real-time human detection and tracking from mobile vehicles, in : IEEE Intelligent Transportation Systems Conference (ITSC'07), 2007, pp. 149–154.

[14] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, MOTChallenge 2015 : Towards a benchmark for multi-target tracking, arXiv :1504.01942 [cs]ArXiv : 1504.01942.

[15] E. Moussy, A. A. Mekonnen, G. Marion, F. Lerasle, A comparative view on exemplar "tracking-by-detection" approaches, in : IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'15), 2015, pp. 1–6.

[16] M. Isard, A. Blake, Icondensation : Unifying low-level and high-level tracking in a stochastic framework, in : Computer Vision – ECCV'98, Springer, 1998, pp. 893–908.

[17] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance : the CLEAR MOT metrics, EURASIP Journal on Image and Video Processing 2008 (2008) 1 :1–1 :10.