Suivi multi-personnes à base de représentations parcimonieuses collaboratives globales

Loïc Fagot-Bouquet, Romaric Audigier, Yoann Dhome, Frédéric Lerasle

Suivi multi-personnes à base de représentations parcimonieuses collaboratives globales

Loïc Fagot-Bouquet[1] Romaric Audigier[1] Yoann Dhome[1] Frédéric Lerasle[2,3]

[1] CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Point Courrier 173, F-91191 Gif-sur-Yvette, France
[2] CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France
[3] Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France

loic.fagot-bouquet@cea.fr

Résumé

Le suivi d'objets reste un sujet complexe et crucial en vision par ordinateur même si plusieurs progrès ont été réalisés au cours des dernières années. Utiliser des représentations parcimonieuses a en particulier été proposé pour définir des modèles d'apparence et appliqué avec succès au suivi mono-objet. Nous considérons dans cet article un algorithme de suivi multi-personnes en ligne utilisant des représentations parcimonieuses afin d'associer les détections aux trajectoires existantes. Nous montrons que les représentations collaboratives sont efficaces dans ce cadre tout en pouvant être rapidement estimées au sein d'un algorithme proche temps réel. Des expérimentations approfondies indiquent que notre approche est compétitive par rapport à d'autres méthodes récentes de suivi et robuste vis-à-vis du détecteur de personnes employé.

Mots Clef

Suivi, parcimonie, représentations collaboratives.

Abstract

Object tracking is still a complex and challenging task in Computer Vision despite a lot of improvements over the past decade. In particular, using sparse representations for designing appearance models has been successfully proposed in single object tracking. In this paper, we consider an online multi-person tracking system using sparse representations to correctly associate detections and existing tracks. We show that collaborative representations are efficient in this context and can be rapidly estimated in a near real-time algorithm. Experiments performed on various datasets show that our approach is competitive against recent tracking methods and robust with respect to the person detector.

Keywords

Tracking, sparsity, collaborative representations.

1 Introduction

Le suivi d'objets est un sujet crucial en vision par ordinateur et est requis dans de nombreuses applications pratiques. Les approches existantes pour aborder ce problème peuvent être regroupées en deux grandes catégories, une regroupant les approches de suivi mono-objet et l'autre celles de type suivi multi-objets. Malgré un grand nombre d'apports dans ces deux sujets, concevoir un algorithme de suivi multi-objets efficace reste encore une tâche complexe du fait des similarités d'apparences et des occultations des cibles.

Le suivi multi-objets peut être réalisé de manière hors ligne [1, 2] en exploitant simultanément toutes les images de la vidéo traitée, ou bien de façon dite en ligne [3, 4, 5, 6] en se limitant aux images passées. Les approches en ligne sont préférables dans les cas où l'aspect temps réel est primordial et donnent néanmoins des résultats assez comparables aux approches hors ligne comme détaillé dans certains articles [7, 8]. Les approches en ligne récentes suivent le plus souvent un fonctionnement de type suivi par détection qui consiste à employer un détecteur de la classe des objets suivis pour estimer les positions des cibles à chaque image. Les détections de l'image en cours de traitement sont ensuite associées aux pistes précédemment estimées afin de reconstruire les trajectoires tout au long de la vidéo. Dans la plupart des cas, des modèles d'apparence spécifiques à chaque cible sont appris en ligne et utilisés pour estimer un score d'affinité d'apparence pour tout couple piste-détection. Certains travaux récents ont alors proposé d'employer des modèles d'apparence plus complexes et robustes afin d'être plus apte à différencier les cibles d'apparence similaire [4, 5, 7, 8]. Nous proposons dans cet article d'employer un modèle d'apparence permettant de mieux discriminer les personnes suivies et de réduire ainsi le nombre de changements d'identité des pistes.

Les performances en suivi mono-objet ont été grandement améliorées au cours des dernières années grâce à l'emploi de modèles d'apparence plus sophistiqués. Les re-

présentations parcimonieuses ont en particulier été utilisées pour construire des modèles d'apparence génératifs et discriminatifs [9, 10, 11]. Bien que le temps d'exécution des premiers algorithmes employant ce type de modèles ne permettait pas d'effectuer un suivi en temps réel, cette limitation peut être traitée par l'utilisation de techniques d'acquisition comprimée comme le montre l'article [12]. Les représentations parcimonieuses ont été récemment employées, en s'inspirant de ces méthodes, en suivi multi-objets pour définir des modèles d'apparence spécifiques [3].

Auparavant, les représentations parcimonieuses ont aussi été employées avec succès en vision par ordinateur dans le domaine de la reconnaissance faciale. L'article [13] a proposé un classificateur à base de représentations parcimonieuses qui considère une représentation collaborative vis-à-vis de l'ensemble des classes concernées et attribue l'exemple représenté à la classe qui présente la plus petite erreur résiduelle. Cette technique a produit des résultats impressionnants au moment de sa publication et a été étendue de diverses manières, en particulier par des méthodes d'apprentissage de dictionnaires. Il a aussi été montré dans [14] que l'aspect collaboratif des représentations jouait aussi un rôle crucial pour expliquer les bonnes performances de cette approche.

Nous proposons dans cet article, ce qui à notre connaissance n'a encore jamais été proposé, d'introduire et d'adapter le concept de représentations parcimonieuses collaboratives entre individus dans le cadre d'un algorithme de suivi multi-personnes en ligne. Nous considérons en effet que des représentations collaboratives permettent de mieux distinguer les individus suivis par rapport à des représentations spécifiques à chaque cible (comme cela a été fait dans [3]) sans employer de caractéristiques élaborées. De plus, elles aident à diminuer le nombre de faux appariements entre les détections et les pistes. Bien que cela nécessite de calculer des répresentations parcimonieuses sur de grands dictionnaires, nous montrons que cette étape peut être réalisée rapidement à l'aide de stratégies d'optimisation à grande échelle en employant des ensembles actifs, comme effectué par exemple dans [15], ce qui permet d'obtenir une méthode proche temps réel.

La structure de cet article est la suivante. Tout d'abord, nous détaillons l'approche proposée dans la section 2, puis expliquons la procédure d'optimisation employée en section 3. La section 4 est ensuite dédiée aux résultats expérimentaux obtenus et la section 5 conclut notre article.

2 Approche proposée

2.1 Aperçu général du système

On définit par $\mathcal{T} = \{T_1, ..., T_k\}$ à l'instant t l'ensemble de toutes les pistes existantes et par $\mathcal{D} = \{d_1, ..., d_l\}$ l'ensemble des détections données par le détecteur de personnes. Chaque détection d est associée à un descripteur $y_d \in \mathbf{R}^m$. Les détections de $\mathcal{D}$ sont soit appariées aux pistes auparavant estimées, grâce à des modèles d'apparence et de mouvement spécifiques, soit utilisées pour démarrer de nouvelle pistes. Les trajectoires qui ont un taux d'association élevé sont considérées confiantes tandis que celles avec un taux d'association faible sont jugées perdues et sont ensuite terminées. La gestion des trajectoires et de leurs statuts est inspirée de [4]. Par soucis de clarté des notations, la dépendance temporelle des variables introduites précédemment et par la suite n'est pas précisée.

L'appariement des détections aux pistes est formulé comme un problème de couplage maximum dans un graphe biparti et est résolu avec un algorithme hongrois ou bien un algorithme glouton. Pour formuler ce problème, une matrice d'affinité $A \in \mathbf{R}^{|\mathcal{T}|\times|\mathcal{D}|}$ est calculée, où A_{ij} est un score d'affinité entre la $i^{ème}$ piste de $\mathcal{T}$ et la $j^{ème}$ détection de $\mathcal{D}$. De façon similaire à [4], les détections sont d'abord associées aux pistes jugées confiantes et appariées au reste des pistes dans un second temps.

Notre approche se distingue des précédentes principalement sur la façon de calculer les scores d'affinité, comme expliqué au sein de la section suivante.

2.2 Scores d'affinité

Le score d'affinité A_{ij} entre T_i et d_j est défini par

$$A_{ij} = \begin{cases} a(i,j) & si \ \ (T_i, d_j) \in \mathcal{L} \\ -\infty & sinon \end{cases}$$

où $a(i,j)$ est un terme orienté sur l'apparence et $\mathcal{L}$ regroupe tous les couples piste-détection (T_i, d_j) qui peuvent être associés ensemble en examinant deux critères : l'un basé sur la distance entre d_j et la position estimée de T_i à l'instant courant et l'autre basé sur leurs formes. $\mathcal{L}$ est défini par

$$\mathcal{L} = \{(T_i, d_j), \ \ dist_{T_i,d_j} < R_i \ \ et \ \ \frac{|h_i - h_j|}{h_i} < S_i\}$$

où $dist_{T_i,d_j}$ est la distance euclidienne entre les positions de T_i et d_j, et h_j (resp. h_i) est la hauteur relative à d_j (resp. T_i). Les valeurs R_i et S_i sont estimées pour chaque piste et augmentent progressivement lorsque l'une d'elle se trouve perdue afin d'autoriser une aire de recherche plus grande.

Les termes $a(i,j)$ sont évalués à l'aide de représentations parcimonieuses. Plus précisément, on appelle $D_i \in \mathbf{R}^{m\times p}$ le dictionnaire relatif à la piste T_i (qui inclut les descriptions des p vues les plus récentes de la personne associée à cette piste). Pour tout ensemble $I = \{i_1, .., i_l\} \in [1..k]$, on appelle $D_I = [D_{i_1}, .., D_{i_l}]$ le dictionnaire commun aux trajectoires $T_{i_1}, ..., T_{i_l}$. Chaque détection d_j est associée à un code $\alpha_{y_{d_j}}$ défini par

$$\alpha_{y_{d_j}} = \underset{\alpha}{argmin} \ [\frac{1}{2}||y_{d_j} - D_{I_{d_j}}\alpha||_2^2 + \lambda||\alpha||_1] \qquad (1)$$

où I_{d_j} est un ensemble d'indices spécifique à d_j et où λ réalise un compromis entre l'erreur de reconstruction

$||y_{d_j} - D_{I_{d_j}}\alpha||_2^2$ et la contrainte de parcimonie $||\alpha||_1$. Le terme $a(i,j)$ est ensuite évalué par

$$a(i,j) = \begin{cases} -\frac{1}{2}||y_{d_j} - D_{I_{d_j}}\alpha^i_{y_{d_j}}||_2^2 & si \;\; i \in I_{d_j} \\ -\infty & sinon \end{cases}$$

où $\alpha^i_{y_{d_j}}$ est déterminé à partir de $\alpha_{y_{d_j}}$ en fixant à zéro toutes les coordonnées non relatives à la piste T_i, et $\frac{1}{2}||y_{d_j} - D_{I_{d_j}}\alpha^i_{y_{d_j}}||_2^2$ est défini comme l'erreur résiduelle de la $i^{ème}$ piste.

2.3 Représentations collaboratives locales et globales

Nous avons considéré deux possibilités pour définir l'ensemble I_{d_j}. La première consiste à définir $I_{d_j} = [1..k]$, ce qui signifie que $\alpha_{y_{d_j}}$ fait intervenir les dictionnaires spécifiques de toutes les trajectoires et est de ce fait appelé representation parcimonieuse collaborative globale (GSC pour *global sparse collaborative representation*). La seconde possibilité est de choisir $I_{d_j} = \{i, (T_i, d_j) \in \mathcal{L}\}$. Cette fois $\alpha_{y_{d_j}}$ est seulement calculé à partir des dictionnaires des pistes pouvant être associées à d_j et cette représentation est donc considérée comme la représentation parcimonieuse collaborative locale de y_{d_j} (LSC pour *local sparse collaborative representation*).
A première vue, il parait inutile de faire intervenir les trajectoires qui ne peuvent être associées à d_j, ceci nécessitant, de plus, d'optimiser (1) sur des dictionnaires plus grands. Cependant, utiliser les représentations locales signifie que les scores d'affinité sont definis en tant qu'erreurs résiduelles trouvées sur des dictionnaires de différentes tailles et il n'est pas évident que ces termes soient réellement comparables.

3 Optimisation

Calculer les représentations globales nécessite de résoudre l'équation (1) sur un dictionnaire pouvant comporter un grand nombre d'éléments car le nombre de trajectoires présentes simultanément peut s'avérer important (par exemple jusqu'à 30 dans certaines vidéos). Les techniques d'optimisation classiques ne peuvent alors être directement appliquées pour réaliser ces optimisations dans un temps acceptable. Nous expliquons dans cette partie comment ce problème peut être efficacement optimisé en employant quelques idées d'optimisation à grande échelle.

3.1 Gradient proximal accéléré

Les méthodes proximales ont été employées pour résoudre le problème (1) et utilisées dans plusieurs domaines d'application, notamment le suivi mono-objet [16]. Ces techniques sont une généralisation des méthodes d'optimisation classiques du premier ordre. Elles sont particulièrement utiles lorsque la fonction objectif à minimiser peut s'écrire sous la forme d'une somme de deux fonctions convexes propres et fermées f et g, avec f différentiable.

Données : D, y, K
Résultat : α_y
$\alpha_y = \emptyset$;
$\mathcal{A} = \emptyset$;
$r = -y$;
répéter
  pour $k = 1$ ***to*** K **faire**
    $i = \underset{j \notin \mathcal{A}}{argmax} \; |D_j^T r|$;
    si $|D_i^T r| > \lambda$ **alors**
      $\mathcal{A} = \mathcal{A} \cup \{i\}$;
    fin
  fin
  $G = D_{\mathcal{A}}^T D_{\mathcal{A}}$;
  En utilisant α_y comme position initiale et G,
  déterminer $\alpha_y = \underset{\alpha}{argmin} \; \frac{1}{2}||y - D_{\mathcal{A}}\alpha||_2^2 + \lambda||\alpha||_1$;
  $r = D_{\mathcal{A}}\alpha_y - y$;
jusqu'à *convergence*;

Algorithme 1 : Stratégie avec ensembles actifs pour accélérer le calcul des représentations parcimonieuses sur de grands dictionnaires.

La fonction objectif de l'équation (1) peut être décomposée de cette manière en choisissant $f(\alpha) = \frac{1}{2}||y - D\alpha||_2^2$ et $g(\alpha) = \lambda||\alpha||_1$.
La méthode de gradient proximale accélérée peut déterminer le minimum de $f + g$ avec une vitesse de convergence en $O(1/i^2)$ où i est l'indice d'itération de la méthode [17]. Cette méthode nécessite de calculer à chaque itération le gradient au point courant $\bigtriangledown f(\alpha_i)$ ainsi que quelques valeurs de l'opérateur proximal de g et quelques évaluations de la fonction f (durant une étape de recherche en ligne). Les étapes limitantes en temps de calcul sont alors le calcul de $\bigtriangledown f(\alpha_i)$ et les évalutations de f, chacune de ces étapes nécessitant $O(nm)$ opérations lorsque le dictionnaire D comporte n éléments de taille m.
Comme expliqué dans [17], il est possible de pré-calculer la matrice de Gram $D^T D$, ce qui nécessite $O(n^2 m)$ opérations, pour pouvoir effectuer chaque itération en $O(n^2)$. Cela peut mener à un gain important en temps de calcul lorsque la condition $n \ll m$ est satisfaite. Dans notre cas, cette condition n'est à priori pas satisfaite et ceci notamment pour les représentations globales. En effet, le nombre d'éléments dans ce dernier cas peut s'avérer important et rend le pré-calcul de la matrice de Gram très coûteux pour un gain faible à chaque itération.

3.2 Stratégie avec ensembles actifs

Quelques techniques ont été proposées afin de traiter le cas de grands dictionnaires. En effet, la condition d'optimalité pour l'équation (1) montre que la solution α vérifie $\alpha_i = 0$ si et seulement si $|D_i^T(D\alpha - y)| \leq \lambda$, où D_i indique la $i^{ème}$ colonne (ou élément) de D. Il est ainsi possible d'employer une stratégie à base d'ensembles actifs (comme fait par exemple dans [15]). Cela consiste à résoudre (1) sur un

sous-ensemble d'éléments de D et à progressivement ajouter de nouveaux éléments à ce sous-ensemble.
Plus précisément, l'équation (1) est optimisée sur un sous-ensemble $\mathcal{A}$ d'éléments de D (ce qui fournit une solution $\alpha_{\mathcal{A}}$). Les éléments de D qui présentent les plus grandes valeurs $|D_i^T(y - D_{\mathcal{A}}\alpha_{\mathcal{A}})|$ au-dessus de λ sont ensuite ajoutés à $\mathcal{A}$. En pratique, seules quelques itérations sont effectuées sur le sous-ensemble $\mathcal{A}$ avant de rajouter de nouveaux éléments. Cette méthode converge vers la solution optimale et il est même possible de pré-calculer la matrice de Gram afin d'accélérer les itérations effectuées sur $\mathcal{A}$ (le nombre d'éléments dans $\mathcal{A}$ étant peu élevé).
Le processus d'optimisation est détaillé dans l'algorithme 1, où K éléments sont ajoutés à $\mathcal{A}$ à chaque étape de sélection.

4 Expérimentations

4.1 Implémentation

Notre méthode a été codée en C++ et testée sur un ordinateur portable sur un seul coeur à 2.7 GHz. Nous avons comparé deux variantes de l'approche proposée, l'une utilisant les représentations parcimonieuses collaboratives globales (GSC) et l'autre les représentations parcimonieuses collaboratives locales (LSC). Nous avons aussi comparé une troisième variante où les termes $a(i,j)$ de la matrice d'affinité sont définis comme étant l'opposé de l'erreur de reconstruction de y_{d_j} à partir du dictionnaire spécifique D_i de la trajectoire T_i. Plus précisement, on considère

$$a(i,j) = -\frac{1}{2}||y_{d_j} - D_i\alpha^i_{y_{d_j}}||_2^2$$

avec

$$\alpha^i_{y_{d_j}} = \underset{\alpha}{argmin}\,[\frac{1}{2}||y_{d_j} - D_i\alpha||_2^2 + \lambda||\alpha||_1]$$

Cette dernière méthode emploie des représentations non collaboratives spécifiques à chaque cible et sera indiquée dans la suite par le sigle TSS (*target specific sparse representation*).
Les positions futures des cibles sont estimées à partir de filtres de Kalman et l'appariement des détections aux pistes, formulé comme un problème de couplage maximum, est résolu avec un algorithme glouton. Nous n'employons pas de caractéristiques élaborées et utilisons directement les valeurs d'intensité RGB des boites (redimensionnées à 30x30 pixels). Tous les paramètres ont été fixés de façon empirique et restent identiques pour toutes les vidéos testées. En particulier, la taille des dictionnaires spécifiques est fixée à 30 éléments et le paramètre λ dans (1) est fixé à 0.1.

4.2 Protocole expérimental

Notre approche a été évaluée sur plusieurs vidéos de types assez variés : PETS S2L1 et S2L2, TownCenter et Parking Lot. Ces vidéos se différencient principalement vis-à-vis du nombre de personnes suivies, du champ de vue et de la fréquence. Comme [4], nous n'utilisons pas d'information de calibration 3D des caméras. Deux détecteurs de personnes différents sont utilisés pour les vidéos de PETS et TownCenter (mêmes jeux de détections que [4]) afin de vérifier la robustesse de notre algorithme par rapport au détecteur employé. Afin de permettre une comparaison valide nous avons comparé notre approche à d'autres algorithmes de suivi en ligne de l'état de l'art [3, 4, 5, 6] sur des vidéos pour lesquelles les détections employées étaient disponibles.
Nous utilisons les métriques CLEARMOT, détaillées dans [18], qui sont évaluées en utilisant le code public de [4] (avec un seuil de recouvrement standard de 0.5 contrairement au seuil inhabituel employé dans [4]). Les métriques suivantes sont prises en compte : l'exactitude du suivi (MOTA), la précision du suivi (MOTP), le nombre de changements d'identité (IDS), le nombre de faux positifs (FP) et le nombre de positions manquantes (MS). Le MOTP considère uniquement la précision de la localisation des individus sans tenir compte des changements d'identité, et le MOTA prend en compte les changements d'identité, les détections manquantes et les faux positifs. Cette dernière métrique est souvent considérée de première importance pour pouvoir juger de la qualité du suivi.
Lorsque celles-ci étaient disponibles, nous avons utilisé les trajectoires fournies par les auteurs des articles comparés [4, 6]. Les résultats pour la vidéo Parking Lot sont par contre ceux indiqués dans les articles des méthodes testées.

Vidéo	**S2L1**		**S2L2**		**Town Center**		**Parking Lot**
Opt. \ Det.	[2]	[4]	[2]	[4]	[4]	[6]	[5]
Naive (fps)	1.9	3.4	0.23	0.44	0.51	0.35	0.92
Ens. actifs (fps)	24	29	5.1	8.7	8.4	6.5	10
Gain	12x	8.5x	22x	19x	16x	18x	10x

TABLE 1 – Temps de calcul (avec détections données).

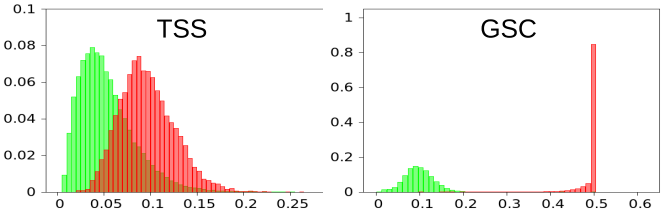

FIGURE 1 – Distribution des coûts d'affinité pour les couples détection-piste corrects (en vert) et incorrects (en rouge) sur la vidéo TownCenter.

4.3 Résultats

Les métriques CLEARMOT pour notre méthode et quelques autres algorithmes sont indiquées dans les tableaux 2 et 3 et quelques exemples de trajectoires obtenues sont présentés en figure 2. Tout d'abord, on peut remarquer

Vidéo	Det.	Méthode	MOTA (%)	IDS	MOTP (%)	FP	MS
S2L1	[2]	TSS	68.8	37	65.5	788	623
		LSC	**70.2**	29	65.4	**737**	**614**
		GSC	69.5	**25**	**65.6**	757	631
	[4]	TSS	71.1	**18**	73.0	451	869
		LSC	71.2	22	72.9	**447**	866
		GSC	**71.3**	19	**73.2**	457	**852**
S2L2	[2]	TSS	38.3	215	65.8	1719	4395
		LSC	40.5	**214**	65.8	1547	4338
		GSC	**41.3**	225	**66.0**	**1502**	**4291**
	[4]	TSS	42.1	214	70.8	1138	4585
		LSC	42.7	210	70.9	1099	4565
		GSC	**43.9**	**194**	**71.1**	**1044**	**4514**
Town Center	[4]	TSS	60.7	211	**71.6**	4275	23580
		LSC	60.6	214	**71.6**	4286	23595
		GSC	**61.3**	**192**	**71.6**	**3983**	**23476**
	[6]	TSS	65.1	225	74.7	7352	17348
		LSC	65.5	210	74.7	7174	**17264**
		GSC	**66.1**	**201**	**74.8**	**6682**	17309
Parking Lot	[5]	TSS	**85.7**	18	**71.3**	276	**759**
		LSC	**85.7**	18	**71.3**	275	761
		GSC	85.6	**17**	**71.3**	**266**	773

TABLE 2 – Métriques CLEARMOT pour les approches proposées (meilleures valeurs en gras et rouge pour le MOTA et les IDS). Seconde colonne : détections employées.

en examinant le tableau 2 que nos approches avec des représentations collaboratives (GSC ou LSC) présentent globalement de meilleures performances en terme de MOTA et de changements d'identité comparées à celle employant des représentations spécifiques par cible (TSS). Comme les individus suivis sont d'apparences assez proches, il semblerait que les dictionnaires spécifiques ne peuvent pas correctement les différencier, comme illustré en figure 1 où nous avons présenté les erreurs de reconstruction des couples piste-détection corrects et incorrects. Bien que l'emploi de caractéristiques plus élaborées pourrait éventuellement améliorer le pouvoir discriminatif des dictionnaires spécifiques, l'emploi de représentations collaboratives améliore déjà les performances de suivi sans recourir à des caractéristiques complexes.

De plus les représentations collaboratives globales (GSC) donnent de meilleurs résultats que ceux obtenus avec les représentations locales (LSC), et nous considérons que ce résultat peut s'expliquer principalement pour deux raisons. D'une part, les représentations globales sont calculées pour une image donnée à partir du même dictionnaire, ce qui évite de comparer des erreurs résiduelles obtenues à partir de dictionnaires de tailles différentes comme cela est fait pour les représentations locales. D'autre part, utiliser des représentations globales semble aussi aider à mieux identifier l'apparition de nouvelles personnes car leurs représentations font intervenir des éléments de trajectoires variées, ce qui se traduit par des erreurs résiduelles plus grandes vis-à-vis des pistes proches.

Vidéo	Det.	Met.	MOTA (%)	IDS	MOTP (%)	FP	MS
S2L1	[2]	[4]	**69.9**	35	**71.2**	805	**557**
		GSC	69.5	**25**	65.6	**757**	631
	[4]	[4]	70.0	21	71.7	543	**827**
		GSC	**71.3**	**19**	**73.2**	**457**	852
S2L2	[2]	[4]	**43.1**	347	**69.5**	**1318**	**4189**
		GSC	41.3	**225**	66.0	1502	4291
	[4]	[4]	39.3	287	69.0	1416	4536
		GSC	**43.9**	**194**	**71.1**	**1044**	**4514**
Town Center	[4]	[4]	60.7	212	71.2	7295	**20549**
		GSC	**61.3**	**192**	**71.6**	**3983**	23476
	[6]	[4]	63.4	446	74.5	9359	16302
		[6]	61.3	318	**80.5**	12309	**14982**
		GSC	**66.1**	**201**	74.8	**6682**	17309
Parking Lot	[5]	[3]*	84.5	**4**	73.2	-	-
		[5]*	79.3	-	**74.1**	-	-
		GSC	**85.6**	17	71.3	266	773

TABLE 3 – Métriques CLEARMOT sur différentes vidéos (meilleures valeurs en gras et rouge pour le MOTA et les IDS). Seconde colonne : détections employées. Le signe * indique que les scores ont été directement repris à partir des articles associés.

Comparée à d'autres approches de l'état de l'art, notre approche avec les représentations globales (GSC) donne le plus souvent de meilleurs résultats en MOTA, comme indiqué dans le tableau 3, et est dans tous les cas toujours comparable avec les autres approches en ligne. Notre méthode produit dans l'ensemble bien moins de changements d'identité (IDS), du fait du pouvoir discriminatif des représentations collaboratives, ce qui traduit un meilleur maintien des trajectoires tout au long du suivi. Concernant le nombre de faux positifs et de positions manquantes, notre approche tend à donner un peu moins de faux positifs mais légérement plus de positions manquantes, comparé à [4]. Sur les vidéos de PETS et TownCenter, l'emploi de détecteurs de personnes différents ne perturbe pas significativement la qualité du suivi obtenu. De ce fait, notre approche est compétitive par rapport aux méthodes actuelles et, de plus, est robuste vis-à-vis du détecteur de personnes employé.

Le temps de calcul de notre approche dépend principalement du nombre de cibles et du nombre de détections par image. Les temps de calcul pour l'optimisation avec la méthode de gradient proximale accélérée et ceux pour la stratégie avec ensembles actifs sont présentés dans le tableau 1 (pour la variante GSC sur un seul coeur). Cela montre que l'utilisation d'ensembles actifs permet de rendre notre approche proche temps réel et que celle-ci pourrait tourner

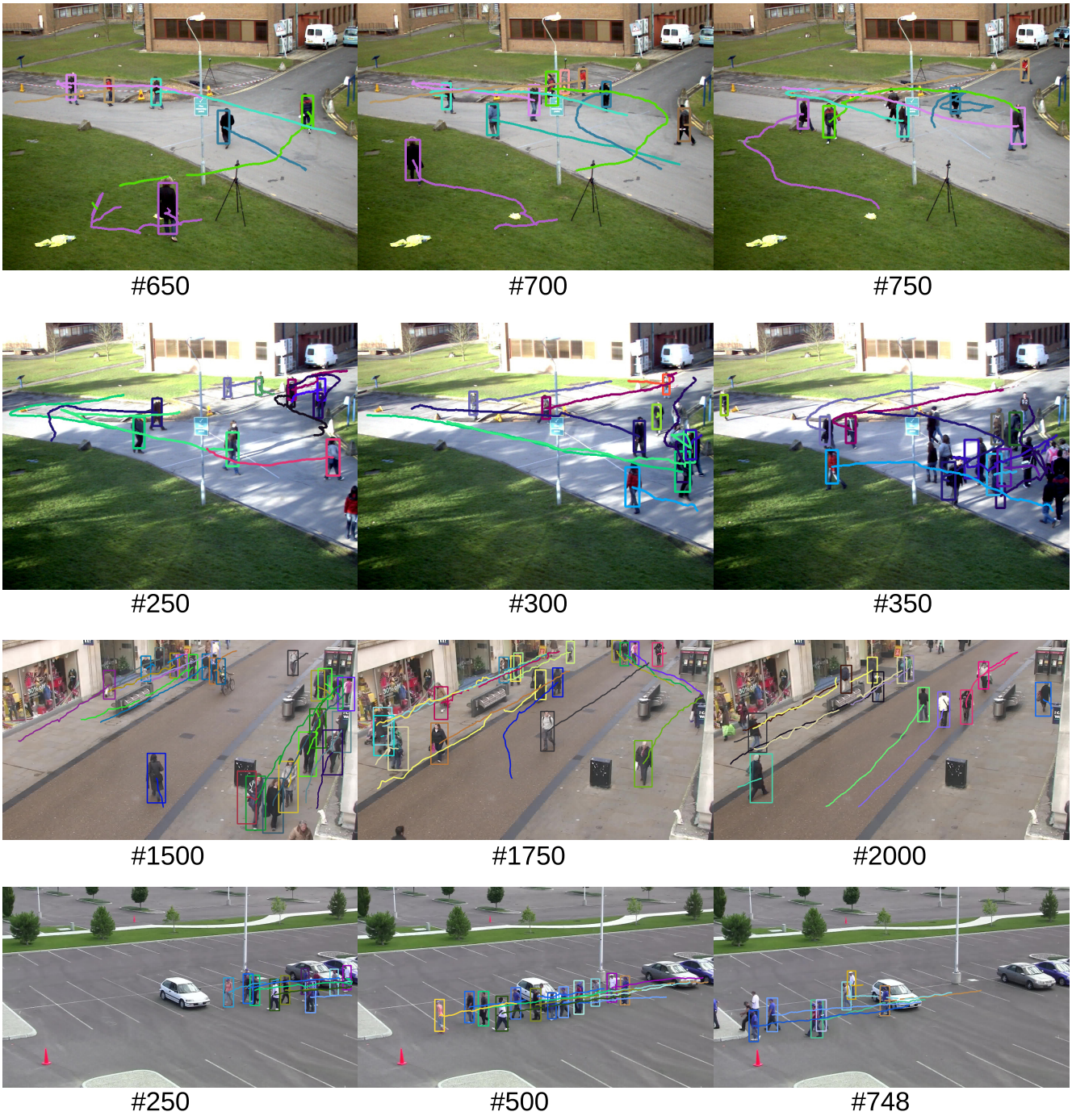

FIGURE 2 – Illustration de nos résultats de suivi pour les vidéos de PETS S2L1, S2L2, TownCenter et Parking Lot (en utilisant les détections de [2] pour PETS, [6] pour TownCenter et [5] pour Parking Lot).

en temps réel en utilisant une implémentation parallèle. On peut noter que le gain en temps de calcul dépend du type de vidéo et du détecteur employé, ceci s'expliquant notamment par le fait que le nombre de trajectoires estimées varie selon les situations.

5 Conclusion

Nous avons proposé dans cet article une méthode de suivi en ligne à partir de représentations parcimonieuses collaboratives globales de toutes les personnes suivies. Nous avons montré que les performances de ces représentations collaboratives dépassent celles obtenues à partir de représentations spécifiques à chaque cible et que ces représentations pouvaient être efficacement calculées malgré le nombre

d'éléments éventuellement important des dictionnaires employés. Une évaluation approfondie de notre système sur plusieurs vidéos montre que cette approche est robuste et compétitive par rapport aux autres méthodes actuelles.
Nous envisageons d'étendre nos travaux en remplaçant l'à priori de parcimonie employé pour calculer les représentations par un à priori plus adapté au contexte du suivi en cours d'estimation. Nous envisageons aussi d'étudier la possibilité de réaliser de façon conjointe l'estimation des représentations parcimonieuses et l'appariement des détections aux trajectoires.

Références

[1] X. Wang, E. Turetken, F. Fleuret, and P. Fua, "Tracking interacting objects optimally using integer programming," *European Conference on Computer Vision (ECCV)*. IEEE, 2014, pp. 17–32.

[2] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2014, vol. 36, pp. 58–72.

[3] M. A. Naiel, M. O. Ahmad, M.N.S. Swamy, Y. Wu, and M. Yang, "Online multi-person tracking via robust collaborative model," *International Conference on Image Processing*. IEEE, 2014.

[4] J. Zhang, L. L. Presti, and S. Sclaroff, "Online multi-person tracking by tracker hierarchy," *Advanced Video and Signal-Based Surveillance*. IEEE, 2012, pp. 379–385.

[5] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1815–1821.

[6] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3457–3464.

[7] S. Bae and K. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1218–1225.

[8] Z. Wu, J. Zhang, and M. Betke, "Online motion agreement tracking," *British Machine Vision Conference*. 2013, BMVA Press.

[9] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2011, vol. 33, pp. 2259–2272.

[10] W. Zhong, H. Lu, and M. Yang, "Robust object tracking via sparsity-based collaborative model," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.

[11] X. Jia, H. Lu, and M. Yang, "Visual tracking via adaptive structural local sparse appearance model," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.

[12] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.

[13] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2009, vol. 31, pp. 210–227.

[14] L. Zhang, M. Yang, and F. Xiangchu, "Sparse representation or collaborative representation : Which helps face recognition ?," *International Conference on Computer Vision*. IEEE, 2011, pp. 471–478.

[15] Y. Mu, J. Wright, and S. Chang, "Accelerated large scale optimization by concomitant hashing," *European Conference on Computer Vision (ECCV)*. IEEE, 2012, pp. 414–427.

[16] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1830–1837.

[17] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, 2013.

[18] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance : The clear mot metrics," *EURASIP J. Image and Video Processing*, 2008, vol. 2008.