Prédiction sélective des traitements pour le suivi d'objet

I. Leang[1] S. Herbin[1] B. Girard[2] J. Droulez[2]

[1] ONERA, Office national d'études et de recherches aérospatiales
[2] ISIR, Institut des Systèmes Intelligents et de Robotique

Chemin de la Hunière et des Joncherettes, 91123 Palaiseau, France
isabelle.leang@onera.fr

Résumé

L'un des problèmes majeurs du suivi d'objet en vidéo est la dérive, c'est-à-dire la mauvaise localisation à un instant donné d'une entité désignée dans une image initiale. Les nombreux algorithmes de suivi visuel ou trackers *sont tous susceptibles de dériver lorsqu'ils sont confrontés à diverses sources de nuisance (occultations, changements d'apparence ou d'illumination, mouvements irréguliers de la cible et de la caméra...) mais sont également caractérisés par des niveaux de robustesse variés à ces phénomènes. L'approche proposée par notre travail est de s'inspirer d'une caractéristique de la vision naturelle, l'attention visuelle, considérée ici comme un ensemble de mécanismes prédictifs de sélection de l'information ou des ressources. Dans le problème étudié, son rôle est alors de combiner dynamiquement les trackers selon leur capacité à assurer globalement le suivi à partir de mesures d'auto-évaluation en ligne de leur comportement et de leur adéquation au contexte. L'effort est porté sur la combinaison de trackers hétérogènes en coût de calcul et en structure fonctionnelle, et sur la capacité de les munir génériquement de telles fonctions d'auto-évaluation.*

Mots Clef

attention visuelle, vision par ordinateur, suivi d'objet, prédiction de dérives, fusion

Abstract

One of the key problems in video object tracking is drift, this is the bad localization of an entity initially appointed in an image at a given moment. The numerous visual tracking algorithms or trackers *are all likely to drift when they are confronted with diverse sources of disturbances (occlusions, appearance changes or illuminations, irregular movements of the target and the camera) but are also characterized by different levels of robustness in response to these phenomena. The approach proposed by our work is inspired by a characteristic of natural vision, visual attention, considered here as a set of predictive mechanisms of selection of the information or the resources. In the studied problem, its role is then to dynamically combine trackers according to their capacity to globally insure tracking from measures of on-line self-assessment of their behaviour and their adequacy to the context. In our combination of heterogeneous trackers, the efforts concern the cost of calculation, the functional structure, and the capacity to equip them generically with such self-assessment functions.*

Keywords

visual attention, computer vision, object tracking, drift prediction, fusion

1 Introduction

Les systèmes naturels de vision possèdent des capacités de robustesse, de polyvalence, d'adaptabilité et d'optimalité des ressources exploitées, qui peuvent être considérées comme globalement supérieures à celles des systèmes artificiels de vision actuels. L'objectif est de s'en inspirer pour améliorer les performances des chaînes de traitements.
Parmi les caractéristiques des systèmes cognitifs naturels, l'exploitation de mécanismes d'attention peut être considérée comme essentielle aux performances des activités perceptuelles et cognitives chez l'homme : "Attention is a core property of all perceptual and cognitive operations" [9]. Plus précisément, nous souhaitons introduire une dimension attentionnelle explicite dans la conception d'algorithmes de "vision par ordinateur" pour en améliorer le compromis performance/coût/adaptabilité.
La perception visuelle est attentionnelle. D'abord au sens d'une "sélection d'information visuelle" qui se manifeste par un mécanisme bien connu, directement observable : les saccades oculaires. La théorie de Treisman et Gelade [20] conjecture que les déclencheurs de ces saccades sont des "indices visuels de la scène qui attirent le regard". Par la suite, ils ont été associés à la notion de "saillance visuelle" [13], et ont donné lieu dans les communautés de robotique et de vision par ordinateur, à la naissance de nombreux modèles de cartes de saillance chargées de représenter en un format synthétique les lieux sur lesquels porte l'attention visuelle [4].
Cependant, lors de la réalisation des tâches de perception visuelle, le cerveau mobilise sélectivement de nombreux

circuits neuronaux (corticaux et/ou sous-corticaux) correspondant à différentes temporalités (courtes ou longues), différentes sous-fonctions cognitives (suivi, reconnaissance, alerte), différents traitements (couleur, forme, mouvement) et combinant des architectures de traitements parallèles et/ou séquentiels. Le phénomène d'attention peut donc être associé à des mécanismes de sélection de traitements cognitifs.
Nous souhaitons donc, dans ces travaux, étendre l'approche courante de modélisation de l'attention visuelle par carte de saillance à une notion plus générale de "sélection de ressources", en distinguant :
– la sélection de l'information visuelle (saccades)
– et la sélection de traitements cognitifs,
pour proposer des solutions capables de gérer de manière optimale un ensemble de capacités cognitives limitées.

Nous voulons appliquer l'attention comme "sélection de ressources" au problème de suivi d'objet ("tracking"), avec comme ressources disponibles des modules de traitement ("trackers"). Les mécanismes d'attention sont assimilables à un problème de sélection et de fusion dynamique de modules réalisant la même fonction de suivi mais avec des caractéristiques différentes, et corrélativement une certaine variété de performances fonctionnelles ou de coûts de calcul. La sélection de traitements consiste alors à combiner le ou les meilleurs traitements à chaque instant pour parvenir à un meilleur résultat de suivi sous contrainte de ressources (mémoire, coûts de calcul, temps d'exécution).
Une caractéristique décisive pour toute stratégie de combinaison de traitements ou fonctions multiples est une capacité d'estimation "en ligne" (en direct) de la qualité des résultats produits, d'une auto-évaluation de leur bon ou mauvais fonctionnement. L'idée générale est d'exploiter ces indicateurs de qualité pour sélectionner et fusionner de manière robuste un certain répertoire de traitements hétérogènes et améliorer globalement les performances tout en contrôlant le coût de calcul. La contribution principale des travaux présentés dans cet article est la description d'une démarche générique permettant de calculer en ligne des indicateurs de bon ou mauvais fonctionnement d'algorithmes de suivi à partir de caractéristiques intrinsèques des trackers.

2 Travaux antérieurs

L'objectif est de construire un tracker capable de localiser un objet par une "boîte englobante" de coordonnées (x_c, y_c, w, h), (x_c, y_c) sont les coordonnées du centre de la boîte et (w, h) sa largeur et sa hauteur. L'idée de base de la fusion de trackers est d'exploiter la complémentarité d'une certaine variété de traitements, de les sélectionner en ligne en fonction de leur utilité, et de les combiner pour produire un meilleur résultat de suivi. L'objectif principal est de contrer les dérives qui peuvent survenir au cours du suivi, venant de l'un ou l'autre des trackers, et de limiter leur impact sur le résultat global.
L'intérêt d'une étape de fusion est de disposer d'une variété de trackers aux comportements complémentaires. Elle peut être provoquée à différents niveaux :
– au niveau des caractéristiques visuelles ;
– au niveau des modèles d'apparence ou de mouvement ;
– au niveau des sorties de traitement pris comme modules séparés ("boîtes noires")

2.1 Fusion de trackers

Fusion par sélection des caractéristiques. Yoon et al. [22] sélectionne un tracker à chaque instant parmi de multiples trackers à filtre particulaire basés chacun sur des caractéristiques discriminantes (HOG, caractéristiques de Haar, intensité). La sélection du meilleur tracker est faite sur la probabilité a posteriori des trackers.
Les travaux de Brasnett et al. [6] et de Penne et al. [17] visent à améliorer le suivi par filtrage particulaire en fusionnant plusieurs caractéristiques hétérogènes (couleurs, textures, contours), et en sélectionnant celles qui mènent à une observation plus discriminatoire.

Fusion par sélection des modèles d'apparence (classifieurs). Zhang et al. [23] utilisent un critère de sélection d'un ensemble de trackers basé sur une minimisation d'entropie. Les trackers diffèrent par leur modèle d'objet (classifieur SVM) pris à des instants différents de la séquence.
Kwon et al. [14] effectuent la sélection d'un tracker à chaque instant parmi un ensemble de trackers échantillonnés dans l'espace des modèles d'apparence, modèles de mouvement, types de représentation d'état et types d'observation. Le tracker choisi est celui qui a la probabilité a posteriori estimée maximale.
Khan et al. [12] utilisent plusieurs modèles de mouvement pour couvrir toutes les trajectoires possibles de la cible.

Fusion par sélection des traitements. La fusion peut s'effectuer sur différents modules de traitements, par exemple la fusion d'un détecteur catégoriel et d'un tracker, ou la fusion de plusieurs trackers différents.
Siebel et al. [19] présentent une fusion de plusieurs modules de traitement pour le suivi de personnes (un détecteur de mouvement, un détecteur de visages, un tracker basé région, un tracker basé forme).
Breitenstein et al. [7] présentent un algorithme de tracking multi-cibles par détection de personnes. Les personnes sont détectées par un détecteur catégoriel de personnes. Un classifieur spécifique à chacune des personnes détectées est entraîné de manière à discriminer les personnes entre elles. Chaque suivi est effectué par filtrage particulaire pour prédire la position de la personne.
Biresaw et al. [3] font la fusion de 2 trackers à filtrage particulaire par une mesure de qualité des prédictions des trackers, pour effectuer la sélection et mettre à jour les modèles d'apparence. Cette mesure de qualité est basée sur l'incertitude spatiale des particules.

Bailer et al. [1] présentent des stratégies de fusion en ligne et hors ligne à partir des sorties de trackers (boîtes englobantes) : vote majoritaire, poids différents pour chaque tracker, optimisation de trajectoire, filtrage des mauvaises sorties avant fusion. Ils montrent que les résultats obtenus par fusion sont meilleurs que ceux des algorithmes de suivi présents dans l'état de l'art.

Notre travail se situe dans la dernière catégorie : combiner un répertoire hétérogène de trackers de performance et de coût variables et de proposer une méthode d'évaluation générique de performance de ces trackers en ligne. L'évaluation de performance permettra d'une part, la sélection dynamique des meilleurs trackers et d'autre part, leur fusion pour améliorer le suivi d'objet.

2.2 Evaluation de la performance en ligne des trackers pour la sélection

Il y a plusieurs manières d'évaluer la performance en ligne des trackers. Une première manière est d'utiliser un modèle de vraisemblance des données conditionnellement aux paramètres du tracker. Une deuxième manière exploite des a priori spatio-temporels pour détecter des comportements aberrants.

Modèle bayésien des trackers. Kwon et al. [14] choisissent le tracker dont la prédiction présente le maximum de vraisemblance par rapport au modèle d'observation de l'objet (toutes les apparences de l'objet depuis le début de la séquence). Cependant, la qualité du modèle d'observation n'est pas contrôlée : il n'y a aucun moyen de savoir si le modèle est toujours bon ou non.
Zhang et al. [23] évitent la dérive du modèle d'observation (sa mise à jour par de mauvais exemples d'apprentissage), en conservant des modèles d'observation à différents instants de la séquence. La sélection du meilleur modèle d'observation se fait sur deux modèles extrinsèques : un modèle d'apparence reliant l'apparence d'un exemple-patch à son label et un modèle spatial reliant le label à sa position dans l'image.
Dans Yoon et al. [22], la probabilité a posteriori que la cible soit dans un état donné sachant une observation, est exprimée comme la somme pondérée des probabilités a posteriori des M trackers. La pondération est la probabilité du tracker (fiabilité) mesurée à partir d'une fonction de vraisemblance de deux modèles d'apparence (apparence récente et apparence reconstruite) de chaque tracker.
Khan et al. [12] utilisent une stratégie de recherche multi-échelles de la cible en intégrant plusieurs modèles de mouvement avec une méthode WLMCMC (Wang-Landau Markov Chain Monte Carlo). Pour T échelles de recherche de l'objet et G modèles de mouvement par échelle, un seul modèle de mouvement est sélectionné à chaque instant : le choix se porte sur la meilleure prédiction d'état du meilleur modèle de mouvement par une mesure du maximum de vraisemblance du modèle d'observation.

A priori spatio-temporels. Chau et al. [8] proposent une méthode d'évaluation en ligne de la performance des trackers (confiance des trajectoires, précision des trackers) par un ensemble de caractéristiques sans utiliser la vérité terrain : longueur de trajectoires avant perte de la cible, zones de dérive, rapport largeur/hauteur de la boîte au cours du temps, aire de la boîte, vitesse de la cible, histogramme de couleurs et sens de déplacement de la cible.
SanMiguel et al. [18] proposent une analyse du changement de comportement du filtre particulaire pour détecter en ligne les dérives du tracker. Ils mesurent l'incertitude du tracker par l'incertitude spatiale de N particules, en analysant les valeurs propres de la matrice de covariance.
Biresaw et al. [2] utilisent un ensemble de trackers par point, chaque point étant associé à un filtre de Kalman. Ils mesurent la qualité de prédiction de chaque tracker en observant les magnitudes de la matrice de covariance du filtre de Kalman. Cette qualité classe les trackers dans deux catégories : les trackers faibles et les trackers forts. Une correction des trackers faibles est effectuée par les trackers forts par une régression par PLS (Partial Least Square).
Zhang et al. [25] entraîne une fonction d'alerte de mauvais fonctionnement des systèmes de vision à partir des sorties (mesure d'erreur ou de précision) couplées aux entrées (image ou caractéristiques extraites) par SVM. Les caractéristiques sont par exemple des points SIFT, des caractéristiques de couleurs, textures, HOG, histogrammes de lignes, LBP, similarités. Il propose ensuite deux métriques d'évaluation des alertes générées par la fonction apprise.

Notre approche est de construire des a priori spatio-temporels sur la sortie des trackers (exploiter les cartes de scores de prédiction) pour détecter leurs mauvais fonctionnements.

3 Sélection et fusion dynamique des trackers

La finalité de la fusion est d'améliorer les performances de suivi. Une étape de sélection est plus que nécessaire pour fusionner strictement les trackers qui amélioreront le suivi. Nous décrirons dans un premier temps, les critères choisis pour évaluer la performance de suivi d'un tracker (section 3.1). Puis dans un second temps, l'étape de sélection par prédiction des dérives (section 3.2). Et dernièrement, l'étape de fusion (section 3.3).

3.1 Critères d'évaluation de performance d'un tracker

Les trackers dont nous disposons ont une structure commune :
– un modèle d'apparence (GMM, SVM, arbre de décision...),

– un ensemble de caractéristiques (Histogrammes de couleurs, HOGs, LBP, Haar...),
– un mécanisme de mise à jour et de sélection d'exemples d'apprentissage,
– une zone de recherche,
– un score de prédiction $s^t(Z^t)$ associé à une proposition de boîte englobante $Z^t = (x_c^t, y_c^t, w^t, h^t)$ à chaque instant t. (x_c^t, y_c^t) sont les coordonnées du centre de la boîte, (w^t, h^t) sont les largeur et hauteur de la boîte à chaque instant. Ce score peut être une confiance, une vraisemblance, une distance.

Les principales difficultés rencontrées lors du suivi telles que les occultations (partielles ou totales), changements d'apparence (couleur, texture, forme, point de vue), mouvements de la caméra (rapidité, flou, déformation), illuminations ; peuvent mener à une perte de la cible.

Précision. L'évaluation de performance des algorithmes de tracking actuels dans un benchmark [21] comprend habituellement la précision de suivi d'un tracker :

– précision par centre de gravité : mesure la distance moyenne des centres de gravité des boîtes englobantes issues de la prédiction du tracker et de la vérité terrain sur une séquence.
– précision par seuil d'erreur de localisation : mesure la proportion d'images pour laquelle la distance (prédiction et vérité terrain) est inférieure à ce seuil, souvent fixé à 20.
– précision par chevauchement : mesure le chevauchement moyen des boîtes englobantes issues de la prédiction du tracker et de la vérité terrain sur une séquence. Le chevauchement de deux boîtes $b1$ et $b2$ est défini comme étant l'intersection des boîtes sur leur union : $\frac{b1 \cap b2}{b1 \cup b2}$.

Robustesse. Une nouvelle mesure a été introduite récemment dans [16], c'est la robustesse de suivi. Elle mesure le nombre de dérives (perte totale de la cible) en moyenne par séquence.

Cette dernière est intéressante car l'un des défis majeurs du suivi d'objet est de réduire le nombre de dérives, et dans le plus favorable des cas, le réduire à 0 (suivi parfait).
Une solution serait d'éviter d'associer des trackers qui dérivent dans notre fusion de trackers. L'idée serait donc d'éliminer les trackers qui dérivent afin de ne garder que ceux qui amélioreront le suivi. Nous proposons de construire une fonction de prédiction en ligne des dérives des trackers en fonctionnement.

3.2 Prédire la dérive d'un tracker

Comme l'a fait Zhang et al. [25], nous présentons une fonction de prédiction des dérives des trackers (anomalies/ruptures de comportement) à partir de leurs sorties (scores de prédiction). Cette fonction permet de sélectionner les bons trackers des mauvais. C'est une méthode qui exploite le comportement intrinsèque du tracker.

Causes de la dérive. Les causes de la dérive peuvent être identifiées à deux niveaux :

– intrinsèque au tracker : le modèle d'apparence n'est plus adapté (accumulation d'erreurs/bruit)
– extrinsèque au tracker : la source (illumination, flou, mauvaise résolution)

Concernant les dérives liées à la source, nous aurions pu nous inspirer des travaux de Zhang et al. pour prédire le mauvais fonctionnement d'un traitement à partir des images brutes d'entrée mais cette perspective n'a pas été abordée.

Détecter les dérives liées au comportement intrinsèque du tracker. La détection de dérives dans un tracker peut-être réalisée à plusieurs niveaux :

– local : en recherchant un "critère-seuil" qui servira à caractériser la fiabilité de la prédiction (score de sortie) du tracker
– spatial : en exploitant non plus un seul score de sortie mais une carte de scores
– **spatio-temporel** : en exploitant la continuité spatio-temporelle de la carte de scores

C'est sur cette dernière (**spatio-temporel**) que se base notre méthode de détection (prédiction) de dérives. La carte de scores contient les scores de sortie du tracker calculés pour toutes les positions de l'image. C'est une carte de probabilités de présence de l'objet en tout point de l'image. Exploiter la continuité spatio-temporelle de ces cartes de scores permet de détecter d'éventuels changements de comportement dans le temps et dans l'espace. L'évolution spatio-temporelle de cette distribution peut être caractérisée par des indices spatio-temporels simples comme l'intensité, la moyenne, la variance, le nombre de maxima locaux. Cette méthode permet de se rendre compte facilement de la précision de prédiction du tracker, uniquement en observant la répartition spatiale des scores autour de la position de la boîte englobante prédite. Nous avons modifié l'outil d'évaluation du challenge VOT [16] pour pouvoir générer des cartes de scores au moment des dérives. Ces cartes de scores serviront par la suite comme base d'apprentissage et d'évaluation pour la prédiction de dérives.
Notons S_t la carte de scores à l'instant t.
$S_t = \{s_t(i,j) | 1 \leq i \leq W, 1 \leq j \leq H\}$, $s_t(i,j)$ est le score en sortie du tracker calculé à la position (i,j) de l'image à l'instant t. W et H sont respectivement la largeur et la hauteur de l'image.
Pour chaque carte S_t, K d'indices de comportement sont extraits. Notons X_t cet ensemble d'indices, $X_t = \{x_t^k | 1 \leq k \leq K\}$. x_t^k est le $k^{\text{ième}}$ indice extrait à l'instant t.
Les indices de comportement extraits peuvent être locaux, c'est-à-dire calculés sur une région de la carte S_t, ou globaux, calculés sur toute la carte.
Nous cherchons à construire la fonction de prédiction des dérives f à partir des indices de comportement extraits de la carte : $f(X_t) = Y_t$ avec $Y_t = \{0, 1\}$.
La fonction prédit 1 pour un bon fonctionnement du tracker, 0 pour une dérive/anomalie de comportement.

3.3 Stratégies de fusion des trackers

La complémentarité des trackers permet d'envisager leur fusion pour améliorer le suivi d'objet (voir Table 2). En effet, dans la plupart des séquences observées de VOT, les trackers ne dérivent pas tous au même moment, nous pouvons donc espérer que leur combinaison assure le suivi complet sur une séquence donnée. Il existe deux modes de fusion des trackers :
– la fusion de trackers sans gestion de coût,
– la fusion de trackers avec gestion de coût.
L'idéal serait de prendre en compte le coût mais cet aspect n'a pas encore été abordé.

Fusion des trackers sans gestion de coût. C'est une stratégie de fusion simple mais coûteuse puisque tous les trackers fonctionnent en parallèle. Tous renvoient une carte de scores. En évaluant chaque sortie de tracker par sa fonction de prédiction de dérives, nous écartons les mauvais trackers. Puis nous fusionnons les sorties des trackers restants.
Nous disposons d'un ensemble de M trackers, noté $T = \{T_m | 1 \leq m \leq M\}$.
Pour chaque image t, chaque tracker T_m renvoie une boîte de sortie $\widetilde{Z}_{m,t}$ et une carte de scores $S_{m,t}$. Pour chaque image t, l'objectif est de fusionner les boîtes des trackers valides pour avoir la meilleure boîte de sortie $\widetilde{Z}_t$, c'est-à-dire la plus proche de celle de la vérité terrain $\widehat{Z}_t$. Nous cherchons donc à minimiser $\parallel \widehat{Z}_t - \widetilde{Z}_t \parallel$. La fusion repose sur les fonctions de prédiction de dérives f_m préalablement calculées pour chaque tracker T_m, capables de prédire une anomalie de comportement du tracker à partir d'indices de comportement extraits des cartes de scores.
Pour chaque carte de scores $S_{m,t}$, nous extrayons K indices de comportement $X_{m,t} = \{x^k_{m,t} | 1 \leq k \leq K\}$.
La boîte issue de la fusion des boîtes de l'ensemble des trackers est : $\widetilde{Z}_t = \sum_{m=1}^{M} \widetilde{Z}_{m,t} * f(X_{m,t})$.

4 Résultats préliminaires pour la prédiction de dérives des trackers

Une modification du toolkit de VOT a permis de générer les cartes de scores des trackers suivants sur 16 séquences RGB annotées (VOT2013), en tout 5681 images :
– NCC Tracker(Normalized Cross Correlation) [15]
– LKT (Lucas Kanade Tracker) [5]
– CT (Compressive Tracking) [24]
– TLD (Track Learn Detect) [11]
– STRUCK (Structured Output Tracking with Kernels) [10]

Chaque séquence comporte la plupart du temps un objet mobile à suivre, souvent placé au centre de la scène par recentrage de la caméra. Plusieurs séquences présentent en revanche un objet fixe mais avec une caméra mobile. Plusieurs difficultés sont rassemblées autour de ces 16 séquences : illuminations, occultations, mouvement de caméra, objets déformables, mouvements d'objet rapides. Le toolkit de VOT est en principe un outil d'évaluation des performances des trackers suivant les critères mentionnés précédemment, la précision et la robustesse. Une des possibilités de cet outil est la réinitialisation automatique des trackers après leur dérivation suivant un critère de chevauchement de boîtes entre la prédiction et la vérité terrain. Le challenge impose une réinitialisation du tracker 5 images après une perte totale de la cible, c'est à dire pour un chevauchement nul des boîtes. Ce qui a permis de générer toutes les fois qu'il y a perte de cible, des cartes de scores au moment des dérives.

Nature des cartes de scores. Les cartes de scores générées à partir des différents trackers sont de natures différentes :
– NCC fournit des scores de corrélation entre $[-1, 1]$
– LKT fournit des nombres d'appariement de points
– CT fournit des rapports de vraisemblance objet/fond
– TLD fournit des scores de confiance par rapport à un modèle objet composé de mini-vignettes
– STRUCK fournit des scores de classification SVM entre $[-1, 1]$

Cependant, malgré l'hétérogénéité des trackers et des cartes de scores (nature, échelle...), les mêmes méthodes d'évaluation des trackers (locale, spatiale, spatio-temporelle) peuvent être appliquées.
En observant qualitativement le changement de comportement des cartes de scores au moment des dérives (voir Figure 1), il semblerait possible d'extraire des indices caractérisant la dérive des trackers. Dans la Figure 1, le changement de comportement est observé globalement sur l'ensemble de la carte de scores mais aussi localement, au niveau de la zone de recherche du tracker. Nous cherchons donc ces indices de comportement locaux et globaux dans ces cartes.

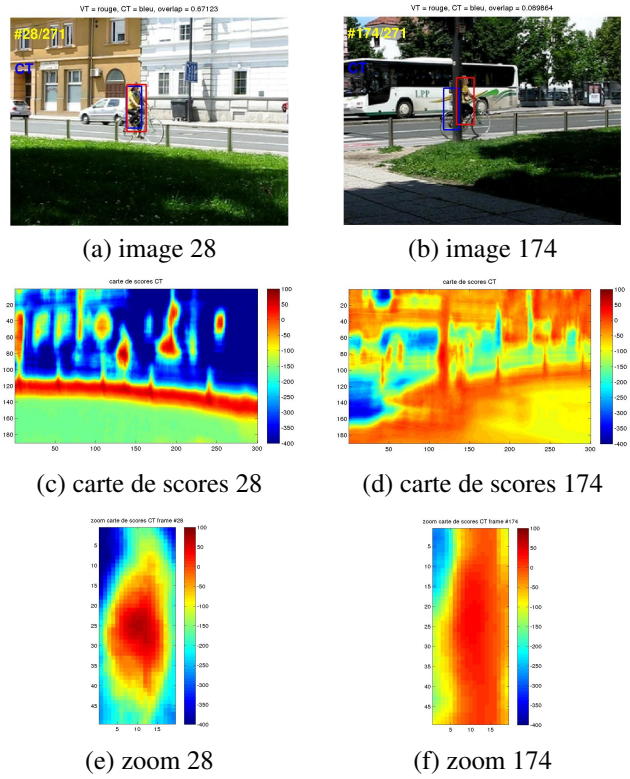
Recherche d'indices de comportement. La recherche d'indices de comportement dans les cartes de scores a été effectué pour 2 trackers, CT et STRUCK. Mais nous pouvons appliquer la même méthode à tous les autres trackers. Pour CT, localement au voisinage de l'objet, l'évolution du nombre de scores à une ligne de niveau fixée sur la fenêtre de recherche (voir Figure 2(a)) donne une indication sur l'étalement de la tâche (variation de taille). Ce critère peut aussi être utilisé globalement, sur toute la carte (voir Figure 2(b)). D'autres indices peuvent être utilisés pour détecter une dissymétrie de la tâche (allongement) et un écrasement en intensité des valeurs centrales. Les Figures 2(a) et 2(b) montrent qu'il est possible de prédire la dérive (signal très piqué juste avant la dérive) avec les deux indices local et global. Nous pouvons envisager de combiner l'indice local et global pour réduire les fausses alarmes, puisqu'ils n'ont pas le même comportement.
Pour STRUCK, 2 indices locaux ont été utilisés pour détecter les changements de comportement : la variance de positionnement des maxima locaux (voir Figure 3(a)) et l'am-

plitude de saut de la position du maximum d'une image à l'autre (voir Figure 3(b)). Une anomalie a lieu lorsque la variance et le saut sont supérieurs à un certain seuil dépendant de la taille de l'objet.
En évaluant nos indices de comportement sur l'ensemble des séquences de VOT, pour CT, on compte 19 bonnes détections de dérives sur 23 dérives du tracker et 91 fausses alarmes. Et pour STRUCK, on compte 12 bonnes détections sur 24 dérives et 25 fausses alarmes (voir Table 1). Malgré un nombre important de fausses alarmes, une majorité des dérives sont bien détectées avec notre jeu d'indices.

Apprentissage des indices de comportement. Plutôt que de rechercher manuellement les seuils de classification des valeurs des indices caractérisant des comportements de dérives, nous voulons généraliser cette classification par apprentissage (SVM). Pour des cartes de scores normales (pas de dérives), les valeurs des indices extraits sont classées dans les exemples positifs. Et pour des cartes de scores anormales (dérives), les valeurs des indices sont classées dans les exemples négatifs. Cependant, nous nous heurtons à une difficulté de labellisation des cartes suivant un critère de chevauchement. Plusieurs raisons peuvent expliquer cet échec :

– le biais engendré par le recentrage de la caméra sur l'objet mobile ne permet pas de s'assurer du bon fonctionnement du tracker, nous constatons souvent après coup que le modèle d'apparence du tracker n'a pas suivi, en observant le comportement de la carte de scores.
– un chevauchement faible entre la boîte prédite et la vérité terrain n'équivaut pas obligatoirement à un changement de comportement drastique de la carte de scores. Notamment lorsque la dérive est lente, le changement de comportement de la carte se produit bien en amont de la perte totale de la cible et ce, de manière progressive. Le tracker se fixe sur un motif du fond et apprend ce motif, qui lentement remplace l'objet. Une fois le motif appris (durée dépendant de l'inertie du modèle d'apparence du tracker), les cartes de scores indiquent un fonctionnement tout à fait normal du tracker d'autant plus que ce motif est invariant au cours du temps.

Dans les conditions de cette base, il n'est pour l'instant pas possible de faire de l'apprentissage à cause des données biaisées.



(a) image 28 (b) image 174

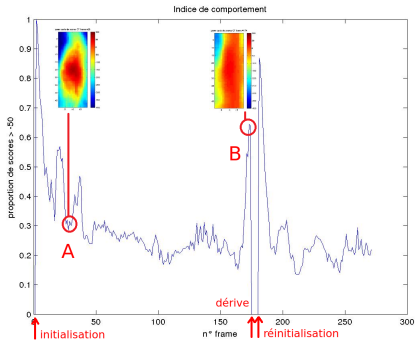
(c) carte de scores 28 (d) carte de scores 174

(e) zoom 28 (f) zoom 174

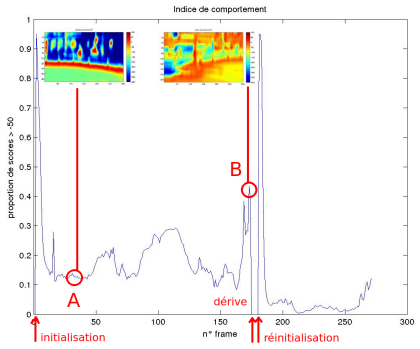
FIGURE 1 – (a) et (b) suivi d'une cycliste (boîtes bleues) pris à deux instants (images 28 et 174) de la séquence bicycle par le tracker CT [24]. Les boîtes rouges correspondent à la vérité terrain. L'image 174 précède la dérive du tracker à cause de l'occultation produite par le poteau. (c) et (d) sont les cartes de scores respectives des images 28 et 174 calculées par le tracker. (e) et (f) sont un zoom de la carte de scores aux positions prédites de la cible par le tracker aux deux instants. (e) les scores d'intensité élevée sont concentrés autour de la position centrale, la prédiction est fiable et précise. (f) le tracker répond de manière uniforme sur toute la fenêtre (même intensité partout), la prédiction est peu fiable.

5 Conclusion et Perspectives

Nous avons présenté dans cet article, une méthode générique d'auto-évaluation de bon ou mauvais fonctionnement des trackers hétérogènes de notre répertoire. Cette méthode exploite la continuité spatio-temporelle des cartes de scores produites en sortie des trackers. Nous pouvons ainsi prédire le disfonctionnement (dérive) des trackers en identifiant des indicateurs de qualité à partir de ces cartes. La prédiction de dérives est une étape essentielle avant la fusion des trackers et permettra d'améliorer considérablement les performances de suivi.
Les travaux futurs viseront à améliorer les performances de prédiction des indicateurs de qualité afin d'augmenter le nombre de bonnes détections de dérives et diminuer le nombre de fausses alarmes. Par ailleurs, une autre piste envisagée serait d'exploiter des indicateurs issus directement de la source. Il pourra être envisagé d'utiliser une plus grande base de données pour faire de l'apprentissage sur ces indicateurs afin de réduire le biais causé par des objets centrés. Notre prochaine étape sera de proposer un chaîne de fusion complète intégrant la méthode générique d'auto-évaluation proposée avec prise en compte du coût de calculs des trackers.



(a) indice local de CT

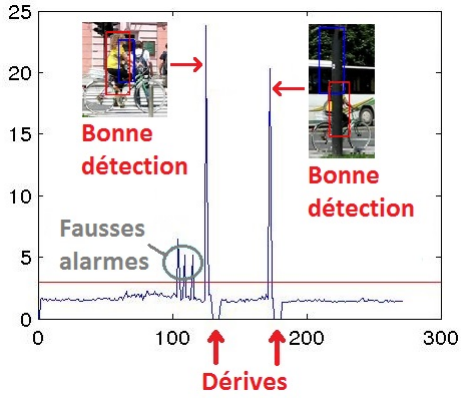


(b) indice global de CT

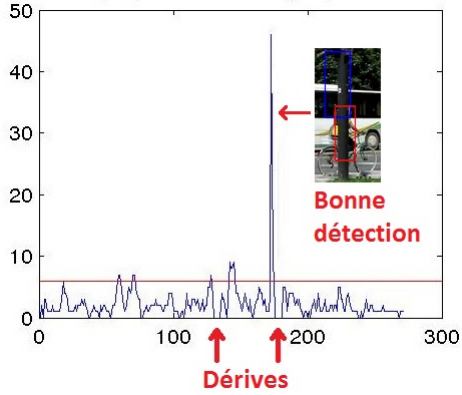
FIGURE 2 – Séquence bicycle (a) évolution d'un indice de comportement de la carte de scores de CT prise localement autour de la cible ("indice local") tout au long de la séquence. Cet indice mesure la proportion de scores supérieure à une ligne de niveau dans la fenêtre. (b) évolution d'un indice de comportement de la carte de scores globale de CT ("indice global") tout au long de la séquence. A l'initialisation et la réinitialisation du tracker, l'indice de comportement est élevé pendant la phase d'adaptation, puis diminue. Les points A et B indiquent deux comportements différents de l'indice, B indique un brusque changement de comportement du tracker, ce qui est en accord avec le comportement du suivi qui est effectivement sur le point de dériver.

	# bonnes détections/ # total détections	# fausses alarmes
CT	19/23 (82%)	91
STRUCK	12/24 (50%)	25

TABLE 1 – Performance des prédicteurs de dérives pour CT et STRUCK (nombre de bonnes détections de dérives sur le nombre total de détections et nombre de fausses alarmes).



(a) indice local 1 de STRUCK



(b) indice local 2 de STRUCK

FIGURE 3 – Séquence bicycle (a) évolution de l'indice local de la carte de scores de STRUCK basé sur la variance de positionnement des maxima locaux. (b) évolution de l'indice local de la carte de scores de STRUCK basé sur l'amplitude de saut du maximum. Dans cette séquence, STRUCK dérive 2 fois. Le premier indice de comportement détecte bien les 2 dérives. Le deuxième indice fait une fausse détection sur la première dérive.

NCC	LKT	STRUCK	TLD
75	99	86	97
97	120	102	114
107	152	155	126
118	175	190	175
128			181
142			
154			
190			

TABLE 2 – Images de la séquence gymnastics pour lesquels les trackers (NCC, LKT, STRUCK, TLD) dérivent. Les trackers sont complémentaires, ils ne dérivent pas aux mêmes moments. Il existe toujours un tracker fonctionnel lorsque tous les autres dérivent.

Références

[1] Christian Bailer, Alain Pagani, and Didier Stricker. A superior tracking approach : Building a strong tracker through fusion. In *Computer Vision–ECCV 2014*, pages 170–185. Springer, 2014.

[2] Tewodros A Biresaw, Andrea Cavallaro, and Carlo S Regazzoni. Correlation-based self-correcting tracking. *Neurocomputing*, 2014.

[3] Tewodros A Biresaw, Andrea Cavallaro, and Carlo S Regazzoni. Tracker-level fusion for robust bayesian visual tracking. 2014.

[4] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1) :185–207, 2013.

[5] Jean-Yves Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 5, 2001.

[6] Paul Brasnett, Lyudmila Mihaylova, David Bull, and Nishan Canagarajah. Sequential monte carlo tracking by fusing multiple cues in video sequences. *Image and Vision Computing*, 25(8) :1217–1227, 2007.

[7] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9) :1820–1833, 2011.

[8] Duc Phu Chau, François Bremond, and Monique Thonnat. Online evaluation of tracking algorithm performance. In *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*, pages 1–6. IET, 2009.

[9] Marvin M Chun, Julie D Golomb, and Nicholas B Turk-Browne. A taxonomy of external and internal attention. *Annual review of psychology*, 62 :73–101, 2011.

[10] Sam Hare, Amir Saffari, and Philip HS Torr. Struck : Structured output tracking with kernels. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 263–270. IEEE, 2011.

[11] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7) :1409–1422, 2012.

[12] Muhammad H Khan, Michel F Valstar, and Tony P Pridmore. A generalized search method for multiple competing hypotheses in visual tracking. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 2245–2250. IEEE, 2014.

[13] Christof Koch and Shimon Ullman. Shifts in selective visual attention : towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer, 1987.

[14] Junseok Kwon and K Lee. Tracking by sampling and integrating multiple trackers. 2013.

[15] JP Lewis. Fast normalized cross-correlation. In *Vision interface*, volume 10, pages 120–123, 1995.

[16] France LIRIS. The visual object tracking vot2014 challenge results.

[17] Thomas Penne, Christophe Tilmant, Thierry Chateau, and Vincent Barra. Mcmc modular ensemble tracking. In *VISAPP (1)*, pages 689–693, 2012.

[18] Juan C SanMiguel, Andrea Cavallaro, and José M Martínez. Adaptive online performance evaluation of video trackers. *Image Processing, IEEE Transactions on*, 21(5) :2812–2823, 2012.

[19] Nils T Siebel and Steve Maybank. Fusion of multiple tracking algorithms for robust people tracking. In *Computer Vision—ECCV 2002*, pages 373–387. Springer, 2002.

[20] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1) :97–136, 1980.

[21] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking : A benchmark. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2411–2418. IEEE, 2013.

[22] Ju Hong Yoon, Du Yong Kim, and Kuk-Jin Yoon. Visual tracking via adaptive tracker selection with multiple features. In *Computer Vision–ECCV 2012*, pages 28–41. Springer, 2012.

[23] Jianming Zhang, Shugao Ma, and Stan Sclaroff. Meem : Robust tracking via multiple experts using entropy minimization. In *Computer Vision–ECCV 2014*, pages 188–203. Springer, 2014.

[24] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang. Real-time compressive tracking. In *Computer Vision–ECCV 2012*, pages 864–877. Springer, 2012.

[25] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems.