Transfert d'apprentissage par un filtre séquentiel de Monte Carlo : application à la spécialisation d'un détecteur de piétons
Transfer Learning by a sequential Monte Carlo filter : application to the specialization of a pedestrian detector

H. Maâmatou[1,2,3] T. Chateau[1] S. Gazzah[2] Y. Goyat[3] N. Essoukri Ben Amara[2]

[1] Institut Pascal UMR 6602 CNRS-UBP-IFMA, Clermont-Ferrand, France
[2] Unité de recherche SAGE, ENISo, Université de Sousse, Tunisie.
[3] Logiroad, Nantes, France

24 Avenue des Landais, BP 80026, 63171 Aubière cedex, France
Houda.Maamatou@etudiant.univ-bpclermont.fr

Résumé

Nous proposons une méthode de transfert d'apprentissage de type transductif basée sur un filtre séquentiel de Monte Carlo pour la spécialisation d'un classifieur générique vers un domaine cible donné. Nous présentons une application de cette méthode pour spécialiser un détecteur de piétons générique à une scène de trafic routier. Les performances enregistrées du détecteur spécialisé sur des données réelles avec un seul faux positif par image, dépassent celles du détecteur générique de plus de 40%.

Mots Clef

Transfert d'apprentissage transductif, Spécialisation, Détecteur de piétons, Filtre de Monte Carlo

Abstract

We propose a method of transductive transfer learning based on a sequential Monte Carlo filter for the specialization of a generic classifier to a target domain. We present an application of this method to specialize a generic pedestrian detector to a traffic scene. The recorded performance of the specialized detector on real data at one false positive per image exceeds that of the generic detector by more than 40% .

Keywords

Transductive transfer learning, Specialization, Pedestrian detector, Monte Carlo Filter

1 Introduction

Un point clé d'un classifieur d'objets générique basé sur l'apparence est la construction d'une base d'apprentissage de grande dimension couvrant toutes les apparences possibles de l'objet dans différentes orientations et un nombre important d'images de négatifs. Bien que les techniques et architectures récentes permettent d'entrainer des classifieurs sur un grand nombre d'exemples, la diversité de certains objets (piétons par exemples) ou de la classe fond conduit à des performances de classifieurs encore trop faibles dans une scène particulière.

Une solution possible à ce problème consiste à utiliser des techniques de transfert d'apprentissage pour spécialiser la classification à une scène cible. Le transfert d'apprentissage se base sur l'utilisation de connaissances acquises antérieurement afin d'améliorer les performances d'une application donnée. Dans [7] deux types principaux de transfert d'apprentissage sont présentés : le type inductif qui suppose la présence de certains exemples étiquetés dans le domaine cible et le type transductif où le domaine cible contient uniquement des données non étiquetées. Ce type d'apprentissage se base sur l'hypothèse que la distribution du domaine source est différente mais reliée à la distribution du domaine cible.

Les méthodes de transfert d'apprentissage se divisent également en deux catégories : une première catégorie modifie un modèle d'apprentissage source pour améliorer son fonctionnement dans un domaine cible et une deuxième catégorie cherche à sélectionner automatiquement les échantillons d'apprentissage qui donnent un meilleur modèle pour la tâche cible. Nous nous intéressons particulièrement au type d'apprentissage transductif et à la deuxième catégorie de méthodes dans un cas de classification d'images. Rosenberg *et al.* [3] utilisent la fonction de décision d'un classifieur basé sur l'apparence d'objet pour sélectionner les exemples d'apprentissage d'une itération à l'autre. Le problème de cette méthode est le paramétrage de la fonction de décision. Si cette dernière est assez sélective alors seulement les données qui sont très semblables seront sélectionnées alors qu'elles peuvent ne pas présenter d'informations importantes de variabilité. Dans un cas contraire, il y a risque d'introduire des données erronées qui dé-

gradent la performance du système dans le temps. Dans le but d'introduire de nouvelles données plus riches en diversité, Levin *et al.* [2] utilisent un système à deux classifieurs indépendants pour collecter des données non étiquetées. Les données étiquetées avec forte confiance par l'un ou l'autre des classifieurs sont ajoutées aux données d'apprentissage pour le ré-entrainement des deux classifieurs. Une autre manière de collecter automatiquement de nouveaux échantillons est d'utiliser une entité externe au système dite "oracle". Un oracle peut être construit à l'aide d'un seul algorithme ou combiner et/ou fusionner plusieurs algorithmes. Dans Nair *et al.*[9] un algorithme à base de soustraction fond-forme est présenté et dans les travaux de Chesnais *et al.* [8] un oracle est construit par trois classifieurs indépendants (apparence, extraction fond/forme et flot optique).

Par ailleurs, certaines solutions concatènent la base source avec les nouveaux échantillons, ce qui augmente la taille de la base au fil des itérations alors que d'autres se limitent uniquement à l'utilisation des nouveaux échantillons perdant ainsi des informations utiles des échantillons sources. Une autre solution intermédiaire a été proposée par Wang *et al.* [5]. Il s'agit de collecter de nouveaux échantillons du domaine cible et de ne sélectionner que des exemples utiles de la base source. Cette méthode est appliquée à la détection de piétons dans une scène de trafic urbain, elle utilise plusieurs de plusieurs indices contextuels tels que le mouvement de piétons, chemin modèle (piétons, voitures,...), localisation, taille et apparence visuelle d'objets pour sélectionner des échantillons positifs et négatifs du domaine cible. Elle se base sur une nouvelle variante des SVM pour sélectionner seulement les échantillons sources bénéfiques à la classification dans la scène cible.

Dans ce papier, nous proposons une formalisation originale de transfert d'apprentissage transductif basé sur un filtre séquentiel de Monte Carlo pour la spécialisation d'un classifieur quelconque à un domaine cible. Ce filtre nous permet de sélectionner les exemples d'une base d'apprentissage qui sont considérés comme des réalisations d'une distribution de probabilité conjointe entre les descripteurs d'échantillons et les classes d'objets. La spécialisation revient alors à estimer la distribution cachée de la base cible. A une itération donnée, un classifieur entrainé sur l'ensemble d'apprentissage de l'itération précédente propose de nouveaux échantillons d'une base cible dont la pertinence est évaluée par une fonction d'observation. Ces derniers sont sélectionnés pour compléter l'ensemble d'apprentissage courant initialisé par une partie propagée de l'ensemble précédent, auquel nous transférons également des exemples d'une base source. Nous avons choisi de garder une taille fixe de la base spécialisée (celle de la base source), pour éviter son augmentation au fil des itérations. La fonction d'observation utilise des informations extraites de la scène cible proches des indices visuels de Wang [5] et de ceux utilisés par Chesnais [8].

Nos principales contributions sont les suivantes :

- Une formalisation originale de transfert d'apprentissage pour la spécialisation d'un classifieur à un domaine donné.
- Une application à la détection de piétons dans des séquences vidéos.
- Une évaluation de la méthode par rapport à l'état de l'art.

Dans la suite, la deuxième section est consacrée à la description de la méthode proposée. La section 3 présente une application de cette méthode dans un cadre de détection de piétons. Nos expérimentations et résultats sont fournis dans la section 4 et nous terminons par une conclusion et quelques perspectives.

2 Spécialisation par un filtre séquentiel de Monte Carlo

Cette section décrit l'idée principale de notre méthode de spécialisation d'un classifieur générique entrainé sur une base source pour obtenir un classifieur spécialisé à une scène particulière (dite scène cible) en proposant un transfert d'apprentissage basé sur un filtre séquentiel de Monte Carlo. La FIGURE 1 présente le synoptique de la spécialisation. Dans la suite, nous appelons détecteur un classifieur qui traite uniquement deux classes (l'objet et l'absence de l'objet).

2.1 Définition du contexte

Notre objectif est de créer un classifieur spécialisé à une scène cible. Pour ce faire, nous proposons de construire itérativement une base spécialisée à la scène cible à partir d'échantillons provenant à la fois d'une base source et de la scène cible qui présente uniquement des échantillons non étiquetés.

Nous supposons que la distribution conjointe des échantillons et des classes cibles (appelée ultérieurement distribution cible) $P(\sum$ échantillons cibles, $\sum$ des classes cibles$)$ peut être approchée par les échantillons labellisés de la base spécialisée. Étant donné que les exemples cibles sont non étiquetés, cette distribution sera estimée par un filtre séquentiel de Monte Carlo à partir d'observations visuelles de la scène.

Nous notons par :

- $\mathcal{D}_k \doteq \{\mathbf{X}_k^{(n)}\}_{n=1,..,N}$ la base spécialisée à l'itération k de taille N, où $\mathbf{X}_k^{(n)} \doteq (\mathbf{x}^{(n)}, y)$ est l'échantillon numéro n avec $\mathbf{x}$ son vecteur de primitives et y son étiquette associée avec $y \in \mathcal{Y}$ (dans un cas de détection $\mathcal{Y} = \{-1; 1\}$, 1 représente l'objet et -1 représente son absence).
- $\Theta_{\mathcal{D}_k}$ un classifieur spécialisé à l'itération k et entrainé sur la base spécialisée construite à l'itération $(k-1)$. Un classifieur associe une étiquette y à un vecteur de primitives $\mathbf{x}$. Nous utilisons un classifieur générique Θ_g à la première itération.

Nous disposons d'une base source $\mathcal{D}^s \doteq \{\mathbf{X}^{s(n)}\}_{n=1,..,N^s}$ de N^s échantillons étiquetés. Nous disposons également d'une base cible $\mathcal{D}^t \doteq \{\mathbf{x}^{t(n)}\}_{n=1,..,N^t}$ qui peut être

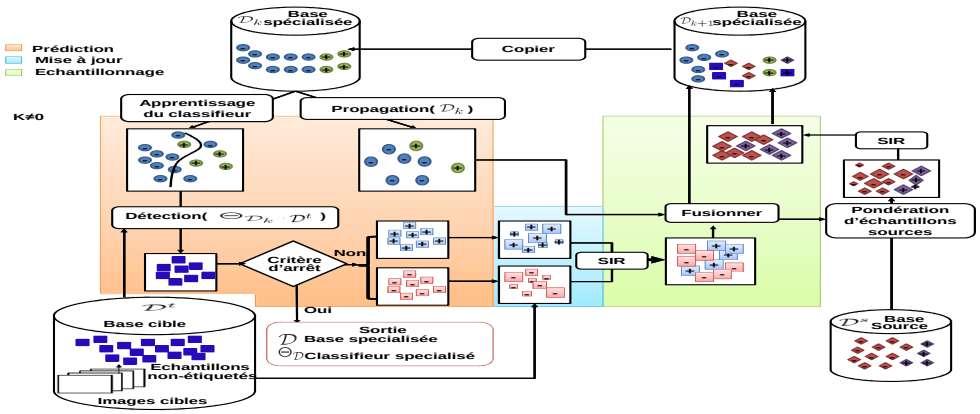


FIGURE 1 – Schéma général de l'approche proposée à une itération donnée ($k \neq 0$)

de grande dimension mais composée uniquement par N^t échantillons non étiquetés qui proviennent d'une vidéo de la scène cible. Cette base est échantillonnée par une stratégie de balayage multi-échelles par fenêtre fixe.

2.2 Filtre séquentiel de Monte Carlo

Notre formalisation repose sur l'hypothèse que la distribution cible est approchée par un ensemble d'échantillons de la base spécialisée. Cependant, ces échantillons initialement inconnus peuvent être déterminés à l'aide d'un processus d'observation issu de la séquence vidéo et des a priori sur la scène cible.
Nous utilisons un filtre séquentiel de Monte Carlo [1] pour estimer cette distribution donc pour sélectionner les échantillons de l'ensemble d'apprentissage. Ceci revient à considérer les échantillons comme étant des particules du filtre et le processus d'observation comme étant la fonction d'observation qui affecte un poids à chaque particule.

Nous notons $\mathbf{X}_k$ un état caché présentant un échantillon inconnu à l'itération k et $\mathbf{Z}_k$ l'observation couplée à $\mathbf{X}_k$ présentant les informations extraites de la séquence vidéo. La méthode séquentielle de Monte Carlo approxime la distribution *a posteriori* $p(\mathbf{X}_k|\mathbf{Z}_k)$ par un jeu de N particules (échantillons dans notre cas) selon l'équation (1) :

$$p(\mathbf{X}_k|\mathbf{Z}_k) \approx \frac{1}{N}\sum_{n=1}^{N}\delta(\mathbf{X}_k^{(n)}) \approx \{\mathbf{X}_k^{(n)}\}_{n=1,..,N} \quad (1)$$

Par la suite, la distribution cible peut être approchée en appliquant de manière itérative l'équation récurrente (2) :

$$p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k+1}) = C.p(\mathbf{Z}_{k+1}|\mathbf{X}_{k+1})\int_{\mathbf{X}_k} p(\mathbf{X}_{k+1}|\mathbf{X}_k)p(\mathbf{X}_k|\mathbf{Z}_{0:k})d\mathbf{X}_k \quad (2)$$

avec $C = 1/p(\mathbf{Z}_{k+1}|\mathbf{Z}_{0:k+1})$. Nous supposons que le processus de récursivité permet de mieux sélectionner les échantillons de la base spécialisée au fil des itérations donc de converger vers la vraie distribution cible et les classifieurs entrainés doivent être de plus en plus performants.

L'équation de récurrence (2) se résout en trois étapes : prédiction, mise à jour et ré-échantillonnage. Ces étapes sont les mêmes que celles d'un filtre à particules utilisé notamment pour la résolution de problèmes de suivi dans le domaine de vision par ordinateur.

Étape de prédiction. L'étape de prédiction consiste à modifier la base spécialisée $\mathcal{D}_k \doteq \{\mathbf{X}_k^{(n)}\}_{n=1,..,N^s}$ selon la dynamique du système $p(\mathbf{X}_{k+1}|\mathbf{X}_k)$ entre deux itérations pour produire l'approximation (3) :

$$p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k}) \approx \{\tilde{\mathbf{X}}_{k+1}^{(n)}\}_{n=1,..,\tilde{N}_{k+1}} \quad (3)$$

Nous notons par $\tilde{\mathcal{D}}_{k+1}$ la base spécialisée estimée à l'itération $(k+1)$ où $\tilde{\mathbf{X}}_{k+1}^{(n)}$ est un échantillon n et $\tilde{N}_{k+1}$ est le nombre d'échantillons proposés par l'étape de prédiction. Cette base est composée de deux sous-ensembles :

1. Un sous-ensemble 1 : c'est un sous-échantillonnage de la base spécialisée précédente, utilisé pour propager la distribution (il s'agit d'un simple tirage aléatoire). Nous imposons le respect d'un ratio entre les différentes classes d'objets (classiquement le même que la base source). Ce sous-ensemble approxime le terme $p(\mathbf{X}_k|\mathbf{Z}_{0:k})$ de l'équation (2) conformément à l'équation (4) :

$$p(\mathbf{X}_k|\mathbf{Z}_{0:k}) \approx \{\mathbf{X}_{k+1}^{*(n)}\}_{n=1,..,N^*} \quad (4)$$

où $\mathbf{X}_{k+1}^{*(n)}$ est l'échantillon n de la base $\mathcal{D}_k$ sélectionné à l'itération $(k+1)$, N^* représente le nombre d'échantillons de ce sous-ensemble avec $N^* = \alpha_t N^s$, $(\alpha_t \in [0,1])$. Le paramètre α_t détermine le nombre d'échantillons à propager à partir de la base précédente.

2. Un sous-ensemble 2 : il s'agit d'une partie des échantillons cibles sélectionnés comme "positifs" par un classifieur entrainé sur $\mathcal{D}_k$ et appliqué sur la base cible $\mathcal{D}^t$. Au cours des premières itérations, ce sous-ensemble contient plusieurs exemples qui ne sont pas de vrais positifs. C'est pourquoi, tous les échantillons

retournés par (5) sont supposés être à la fois positifs et négatifs et la décision entre les deux étiquettes est assurée par la fonction d'observation dans l'étape de mise à jour.

$$\{\breve{\mathbf{X}}_{k+1}^{(n)}\}_{n=1,..,\breve{N}} \doteq \{(\mathbf{x}^{(n)}, y)\}_{y\in\mathcal{Y}\ ;\mathbf{x}^{(n)}\in\mathcal{D}^t/\Theta_{\mathcal{D}_k}(\mathbf{x}^{(n)})>0} \quad (5)$$

$\breve{\mathbf{X}}_{k+1}^{(n)}$ représente l'échantillon cible n proposé pour la base de l'itération $(k+1)$ sachant qu'il est classé positif à l'itération k

Dans ce qui suit, nous notons respectivement les fonctions retournant ces deux sous-ensembles par $Propagate(\mathcal{D}_k)$ et $Detect(\Theta_{\mathcal{D}_k}, \mathcal{D}^t)$.

Étape de mise à jour. Cette étape affecte un poids $\breve{\pi}_{k+1}^{(n)}$ à chaque échantillon $\breve{\mathbf{X}}_{k+1}^{(n)}$ retourné par $Detect(\Theta_{\mathcal{D}_k}, \mathcal{D}^t)$ pour définir le terme de vraisemblance (6) tout en utilisant une fonction d'observation.

$$p\left(\mathbf{Z}_{k+1}|\mathbf{X}_{k+1} = \breve{\mathbf{X}}_{k+1}^{(n)}\right) \propto \breve{\pi}_{k+1}^{(n)} \quad (6)$$

La fonction d'observation utilise des indices contextuels visuels et des a priori extraits à partir de la séquence vidéo de la scène cible. Parmi les informations utilisées largement dans l'état de l'art nous citons le mouvement d'objets, la soustraction de fond et/ou chemin modèle pour chaque objet. Nous supposons que le poids affecté inclut une information sur la classe d'appartenance de l'échantillon. La sortie de cette étape est un ensemble d'échantillons cibles pondérés que nous appelons base cible pondérée (7) par la suite :

$$\{(\breve{\mathbf{X}}_{k+1}^{(n)}, \breve{\pi}_{k+1}^{(n)})\}_{n=1,..,\breve{N}_{k+1}} \quad (7)$$

avec $(\breve{\mathbf{X}}_{k+1}^{(n)}, \breve{\pi}_{k+1}^{(n)})$ le couple formé d'un échantillon cible et son poids associé et $\breve{N}_{k+1}$ représente le nombre des échantillons qui ont un poids différent de zéro.

Étape de re-échantillonnage. La base cible pondérée approxime la distribution conditionnelle $p(\breve{\mathbf{X}}_{k+1}|\mathbf{Z}_{k+1})$ des exemples cibles donnés par les observations à l'itération $(k+1)$. Nous utilisons l'algorithme SIR (Sampling Importance resampling)[1] dans le but de générer un nouvel ensemble d'échantillons cibles non pondérés estimant la même distribution (8) :

$$p(\breve{\mathbf{X}}_{k+1}|\mathbf{Z}_{k+1}) \approx \{\breve{\mathbf{X}}_{k+1}^{*(n)}\}_{n=1,..,\breve{N}_{k+1}^*} \quad (8)$$

où $\breve{\mathbf{X}}_{k+1}^{*(n)}$ est l'échantillon n sélectionné à l'itération $(k+1)$ à partir de la base cible pondérée. A ce niveau, la distribution *a posteriori* $p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k+1})$ est approchée par (9) :

$$p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k+1}) \approx \{\mathbf{X}_{k+1}^{*(n)}\}_{n=1,..,N^*} \cup \{\breve{\mathbf{X}}_{k+1}^{*(n)}\}_{n=1,..,\breve{N}_{k+1}^*} \quad (9)$$

Dans le but d'utiliser la distribution source pour améliorer l'estimation de la distribution cible, nous proposons de compléter la nouvelle base spécialisée par un transfert des échantillons sources sans modifier la distribution *a posteriori*. Notre idée consiste à estimer pour chaque échantillon source, sa probabilité pour qu'il appartienne à $p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k+1})$ en utilisant un algorithme non paramétrique (KDE ou estimateur KNN). En utilisant ces probabilités, l'algorithme SIR sélectionne des échantillons sources qui approximent $p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k+1})$ selon l'équation (10) :

$$p(\mathbf{X}_{k+1}|\mathbf{Z}_{0:k+1}) \approx \{\mathbf{X}_{k+1}^{s*(n)}\}_{n=1,..,\breve{N}_{k+1}^{s*}} \quad (10)$$

où $\mathbf{X}_{k+1}^{s*(n)}$ est l'échantillon n de la base source sélectionné à l'itération $(k+1)$. $\breve{N}_{k+1}^{s*}$ représente le nombre d'échantillons sources sélectionnés. Ce nombre est déterminé à l'aide de l'équation (11) :

$$\breve{N}_{k+1}^{s*} = N^s - (N^* + \breve{N}_{k+1}^*) \quad (11)$$

Finalement, la nouvelle base à l'itération $(k+1)$ est construite par l'union des trois ensembles (12) :

$$\mathcal{D}_{k+1} \doteq \{\mathbf{X}_{k+1}^{*(n)}\}_{n=1,..,N^*} \cup \{\breve{\mathbf{X}}_{k+1}^{*(n)}\}_{n=1,..,\breve{N}_{k+1}^*} \cup \{\mathbf{X}_{k+1}^{s*(n)}\}_{n=1,..,\breve{N}_{k+1}^{s*}} \quad (12)$$

Le Tableau1 présente des définitions de certaines notations utilisées dans le reste du papier et Algorithme 1 résume notre méthode de spécialisation par transfert d'apprentissage basée sur un filtre séquentiel de Monte Carlo.

3 Application de l'approche de spécialisation à la détection de piétons

Cette section présente l'application de la méthode de spécialisation dans un cas de détection de piétons à partir d'une vidéo provenant d'une caméra statique.

Dans une première étape d'initialisation nous avons appliqué une méthode d'extraction fond/forme sur la vidéo cible pour extraire M images pour la spécialisation tout en respectant un pas fixe. Ensuite, nous avons sélectionné les blobs associés à ces images.

Dans ce qui suit, nous présentons les trois étapes du filtre.

3.1 Étape de prédiction

L'étape de prédiction consiste à sélectionner aléatoirement un sous-ensemble de la base spécialisée de l'itération précédente (sous-ensemble 1) d'une part et d'autre part consiste à proposer une liste de détections (sous-ensemble 2) fournie par un classifieur appliqué sur la base cible.

Cette dernière est formée en appliquant une stratégie de balayage par fenêtre glissante à plusieurs échelles sur toutes les M images extraites pendant l'étape d'initialisation. Une fenêtre est considérée détection si son score est supérieur à un seuil donné (ce seuil égal zéro dans notre cas). Cette stratégie permet d'explorer des exemples positifs et des exemples faux positifs.

Tableau 1 – Définition de certaines notations utilisées dans ce papier

- $Learn(\Theta, \mathcal{D})$: entraine un classifieur Θ sur la base $\mathcal{D}$
- Obs_fn : Fonction d'observation
-$W(\mathcal{D}^s)$: pondération des échantillons de la base source
- $SIR(\{\pi^{(n)}\}_{n=1,..,N})$: applique un échantillonnage SIR
- $compute\_overlap(P)$: calcule le "overlap_score" de P
- $compute\_accumulation(P, L, M, w, h)$: calcule le "accumulation_score" de P
- $Area(D)$: calcule la surface du rectangle D

FIGURE 2 – Résultat de l'étape de prédiction :(a) Détections du détecteur générique appliqué sur les images de la base CUHK_Square. Résultat de l'étape de mise à jour : (b) Exemples classés positifs, (c) Exemples classés négatifs. Les échantillons englobés par un rectangle rouge indiquent un échec de classification

Ensuite, une étape de mean-shift est appliquée pour fusionner toutes les fenêtres autour d'une même détection (FIGURE 2.(a)). Pour chaque détection, nous proposons à la fois un échantillon positif et un négatif.

Pour la première itération, nous avons uniquement le (sous-ensemble 2) parce que la base spécialisée est vide. Nous avons utilisé un détecteur générique entrainé sur la base INRIA Person Dataset [6] similaire à celui proposé par Dalal *et al.* dans [6]. La base INRIA contient deux classes : une classe "positive" de 2416 imagettes de piétons de taille (64 x 128) et une classe "négative" de 1218 images qui ne contiennent pas de piétons et que nous utilisons pour générer 12000 imagettes de même taille que celles de la classe positive. Dans ce travail, nous avons utilisé le descripteur HOG comme vecteur de primitives et l'entrainement des classifieurs générique et spécialisé se fait avec l'implémentation SVMLight [1].

3.2 Étape de mise à jour

Cette étape sert à attribuer des poids aux échantillons de "sous ensemble 2" fournit par la prédiction en se basant sur une fonction d'observation. Cette dernière calcule certains indices contextuels et après selon l'étiquette de la proposition et les valeurs des indices calculés, attribue un poids (Algorithme 2 décrit les détails de la fonction d'observation).

1. http ://svmlight.joachims.org

Algorithme 1 Transfert d'apprentissage pour la spécialisation

Entrée: Base source $\mathcal{D}^s$
Classifieur générique Θ_g
Vidéo d'une scène cible et la base $\mathcal{D}^t$ associée
Nombre des échantillons sources N^s.
Paramètres α_t, α_s.
Sortie: La dernière base spécialisée $\mathcal{D}_k$
Le dernier classifieur $\Theta_{\mathcal{D}_k}$
$k \leftarrow 0$
$stop \leftarrow faux$
Tant que $stop \neq vrai$ **faire**
{Étape de prédiction}
Si $(k = 0)$ **alors**
$N^* \leftarrow 0$
$\tilde{\mathcal{D}}_{k+1} \leftarrow Detect(\Theta_g, \mathcal{D}^t)$
Sinon
Learn $(\Theta_{\mathcal{D}_k}, \mathcal{D}_k)$
$N^* \leftarrow \alpha_t N^s$
$\{\mathbf{X}^{*(n)}_{k+1}\}_{n=1,..,N^*} \leftarrow Propagate(\mathcal{D}_k)$
$\tilde{\mathcal{D}}_{k+1} \leftarrow Propagate(\mathcal{D}_k) \cup Detect(\Theta_{\mathcal{D}_k}, \mathcal{D}^t)$
FinSi
Si $(|\tilde{\mathcal{D}}_{k+1}|/|\tilde{\mathcal{D}}_k| >= \alpha_s)$ **alors**
$stop \leftarrow vrai$
Sinon
{Étape de mise à jour}
$\{(\breve{\mathbf{X}}^{(n)}_{k+1}, \breve{\pi}^{(n)}_{k+1})\}_{n=1,..,\breve{N}_{k+1}} \leftarrow Obs\_fn$
{Étape de re-échantillonnage}
$\{\breve{\mathbf{X}}^{*(n)}_{k+1}\}_{n=1,..,\breve{N}^*_{k+1}} \leftarrow SIR(\{\breve{\pi}^{(n)}_{k+1}\}_{n=1,..,\breve{N}_{k+1}})$
$\{(\mathbf{X}^{s(n)}_{k+1}, \breve{\pi}^{s(n)}_{k+1})\}_{n=1,..,N^s} \leftarrow W(\mathcal{D}^s)$ {Pondération des échantillons sources}
$\breve{N}^{s*}_{k+1} \leftarrow N^s - (N^* + \breve{N}^*_{k+1})$
$\{\mathbf{X}^{s*(n)}_{k+1}\}_{n=1,..,\breve{N}^{s*}_{k+1}} \leftarrow SIR(\{\breve{\pi}^{s(n)}_{k+1}\}_{n=1,..,N^s})$
$\mathcal{D}_{k+1} \leftarrow$équation (12)
FinSi
$k \leftarrow k + 1$
FinTQ

Indices contextuels. Dans le but de déterminer les propositions correctes, nous calculons pour chaque proposition deux scores qui sont le "overlap_score" et le "accumulation_score". Le "overlap_score" décrit le rapport entre la surface d'intersection et la surface d'union de deux rectangles, où le premier P présente la proposition et l'autre B présente le blob fournit par la méthode d'extraction fond/forme. Nous notons ce score par λ_o, ($\lambda_o \in [0, 1]$) dont le calcul se fait conformément à l'équation (13)

$$\lambda_o = 2(Area(D\&B))/(Area(D) + (Area(B)) \quad (13)$$

Le "accumulation_score" (noté par λ_a, $\lambda_a \in [0, 1]$) dénote un taux de présence de détections dans une même position de l'image sur l'ensemble M de spécialisation. Ce score met en évidence les zones du fond de l'image pour lesquelles le classifieur répond positivement.

Algorithme 2 Fonction d'observation pour pondérer les échantillons cibles

Entrée: M Nombre d'images pour la spécialisation
Dimension d'une image(w,h)
Liste L des propositions cibles positives et négatives
α_p seuil pour affecter une pondération positive
Sortie: Liste L des propositions mis à jour

```
Pour i = 1 to L faire
  {calcul des indices contextuels }
  P(i).λ_o ← compute_overlap(P(i))
  P(i).λ_a ← compute_accumulation(P(i), L, M, w, h)
  {Affectation des poids }
  Si (P(i).label = 1) alors

    Si (P(i).lamda_o ⩾ α_p) alors
      P(i).pi ← P(i).lamda_o
    FinSi
  Sinon
    Si ((P(i).lamda_o = 0.0)&(P(i).lamda_a > 0.0) alors
      P(i).pi ← P(i).lamda_a
    FinSi
  FinSi
FinPour
```

Pondération des échantillons cibles. Dans notre travail, nous supposons qu'une proposition positive peut être vraie si son "overlap_score" est élevé et par la suite la fonction d'observation doit attribuer un poids important à une telle proposition (FIGURE 2.(b)). Un "overlap_score" est classé élevé s'il dépasse un seuil α_p, où α_p un seuil fixe déterminé empiriquement. Ceci réduit le nombre de propositions positives. Pour remédier à ce problème, nous appliquons une réflexion horizontale à chaque échantillon classé positif pour augmenter le nombre et introduire de la variabilité au sein des échantillons.

D'un autre coté, le "accumulation_score" nous aide à prendre la décision à propos une proposition négative. Une proposition négative qui a un λ_a élevé est probablement une proposition vraie et ceci signifie qu'elle doit avoir un poids important par la fonction d'observation. Les propositions négatives (FIGURE 2.(c)) sont appelées "exemples-difficiles" parce qu'elles ont été classées comme piétons par le classifieur de l'itération précédente. Utiliser ces propositions dans l'ensemble d'échantillons négatifs de la base spécialisée de l'itération suivante, peut améliorer la performance du détecteur spécialisé.

3.3 Étape de re-échantillonnage

Il s'agit de sélectionner les exemples cibles et source pour construire la base spécialisée.

Sélection des exemples cibles. L'étape de mise à jour fournit une base cible pondérée qui approche la distribution cible par les échantillons eux-mêmes. Il est possible que cette base contienne un objet ou une partie d'objet mobile non-piéton classé comme piéton parce que $\lambda_o \geqslant \alpha_p$. Il est possible aussi de classer un piéton statique dans la liste des négatifs parce que son "overlap_score" est nul et son "accumulation_score" est assez élevé.
Pour ne pas insérer ce type d'échantillons dans la base de l'itération suivante, nous évitons de re-sélectionner tout échantillon déjà présent dans l'ensemble propagé de la base précédente. Nous appliquons l'algorithme SIR pour construire une base cible non-pondérée qui a le même nombre d'échantillons que la base pondérée.

Sélection des exemples sources. La FIGURE 2 montre que la base cible calculée automatiquement peut présenter certains échantillons avec de fausses étiquettes. De plus, elle est insuffisante pour générer un détecteur performant pour la détection dans la scène cible. Cependant, la base source contient des échantillons étiquetés dont plusieurs sont très semblables aux échantillons cibles et qui peuvent être bénéfiques pour la spécialisation du détecteur.
Dans l'objectif de transférer à la base spécialisée des échantillons de la base source sans modifier sa distribution, nous affectons à chaque exemple de la base source un poids $\breve{\pi}_{k+1}^{s(n)}$. Ce poids définit la probabilité que cet échantillon soit issu de la base cible. Dans ce travail, cette probabilité est estimée de manière non paramétrique par une méthode de type KPPV (approximée par l'interface FLANN [2] de la librairie OpenCV [3]).
En d'autres termes, la base source contient des exemples de piétons pris avec plusieurs points de vues et orientations. Cependant, une scène de vidéo surveillance capturée par une caméra statique présente des piétons qui sont pris du même point de vue et avec un nombre très réduit d'orientations. Les négatifs sources peuvent parvenir de plusieurs domaines, par contre le fond d'une scène de vidéo surveillance est le même et il est presque statique. Tenant compte de ces constations, nous supposons que si notre scène présente des piétons en vue de face, le processus de pondération affecte un poids fort aux exemples sources pris en vue de face et un poids faible aux autres échantillons. De même, les sources négatives appartenant aux mêmes catégories que celles du fond de la scène ont un poids plus important que les autres. Ceci nous permet d'approcher la distribution cible avec des échantillons sources similaires visuellement à ceux de la scène cible. Nous calculons le nombre d'exemples à sélectionner selon l'équation (11) et nous faisons appel de nouveau à l'algorithme SIR pour sélectionner des échantillons sources et les inclure dans la base spécialisée. Comme les poids utilisés dans l'algorithme SIR sont proportionnels à la probabilité qu'un échantillon soit issu de la distribution cible, la distribution générée en sortie de l'échantillonnage est théoriquement la même que la distribution cible.
A la fin de cette étape, nous concaténons les trois sous-ensembles : sélection aléatoire de la base précédente, base cible et source non-pondérées pour former la base spéciali-

2. http ://www.cs.ubc.ca/research/flann/
3. http ://opencv.org/

sée. Le processus de spécialisation s'arrête lorsque le rapport entre la taille de la base estimée à l'itération k et celle de la base estimée à l'itération $(k-1)$ dépasse un seuil α_s préalablement déterminé ($\alpha_s = 0.80$ fixé empiriquement dans notre cas). Une fois la spécialisation terminée, le détecteur obtenu peut être utilisé pour la détection de piétons dans la scène cible en se basant uniquement sur leurs apparences.

4 Expérimentations et Résultats

Cette section présente les différents tests réalisés pour évaluer les performances de la spécialisation.
Nous avons testé notre méthode sur la base CUHK_Square [5], c'est une séquence vidéo de trafic routier qui dure 60 minutes. Pour la spécialisation, nous avons utilisé 352 images extraites uniformément de la première moitié de la vidéo et pour le test 100 images sont extraites de la deuxième moitié. Nous avons utilisé la vérité terrain fournie par Wang *et al.* [5] et la règle PASCAL [4] (la détection est considérée bonne si le recouvrement avec la vérité terrain dépasse 0.5) pour calculer le taux de vrais positifs.

La durée moyenne d'une itération de spécialisation est environ une heure sur une machine Intel(R) Core(TM) i7-3630QM 2.4G CPU pour traiter une base de 14416 échantillons et appliquer des fonctions d'observations sur 352 images de taille (1440×1152). Une optimisation de ce temps est possible mais nous n'avons pas encore travaillé sur l'amélioration de nos programmes.

4.1 Influence du paramètre α_t

Le paramètre α_t est utilisé pour ajuster le nombre d'échantillons à propager d'une itération à une autre.
Le Tableau 2 montre les performances des détecteurs pour différentes valeur de α_t. Il présente un maximum pour une valeur de α_t égale à 0.75.

Tableau 2 – Performance de détection de différents détecteurs pour différentes valeurs du paramètre α_t à un seul faux positif par image

α_t	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Performance en %	67,2	70,2	73,9	**77**	75,38

4.2 Convergence de la spécialisation

La FIGURE 3.(a) compare la performance du détecteur spécialisé pour plusieurs itérations avec celle du détecteur générique. Dès la première itération, la performance du détecteur spécialisé dépasse celle du détecteur générique de plus de 30% pour un faux positif par image. Les courbes montrent que la spécialisation converge au bout de quatre itérations avec un taux de vrais positifs supérieur à 70%. La FIGURE 3.(b) présente le nombre de détections en fonction du numéro d'itération. Nous remarquons que le nombre des détections se stabilise à partir de l'itération 4 (c'est l'itération d'arrêt selon le critère α_s défini précédemment).

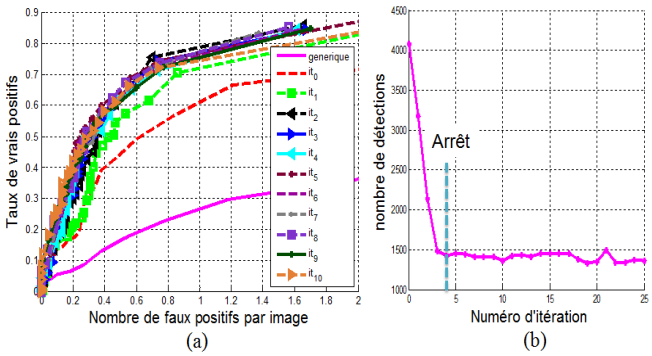


FIGURE 3 – Convergence de la spécialisation : (a) Performance du détecteur générique et des détecteurs spécialisés à plusieurs itérations, (b) Nombre de détections au fil des itérations, Convergence au bout de l'itération 4

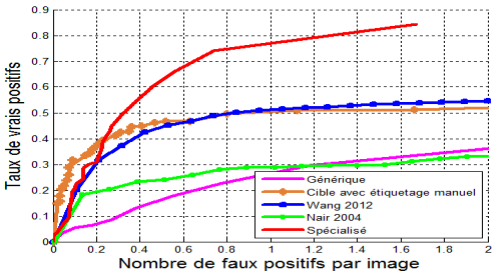
Étant donnée qu'il est difficile de calculer théoriquement la distribution cible, nous avons utilisé la Divergence Kullback-Leiber (D_KL) qui permet de calculer l'écart entre une distribution empirique et une loi (ou distribution théorique) pour mesurer l'écart entre les échantillons positifs des bases spécialisées de quatre premieres itérations et les vrais échantillons positifs étiquetés manuellement de tout l'ensemble M des images de spécialisation.
Le Tableau 3 montre que la Divergence Kullback-Leiber est en diminution ce qui nous permet d'expliquer d'une autre manière la convergence de la spécialisation vers la vraie distribution cible.

Tableau 3 – D_KL entre les échantillons des bases spécialisées au fil des itérations et les échantillons étiquetés manuellement.

Itération	**0**	**1**	**2**	**3**
D_KL	430.97	351.97	306.47	276,05

4.3 Comparaison de la performance avec des méthodes d'état de l'art

Nous comparons la performance globale de notre méthode avec les méthodes de l'état de l'art suivantes :
- Détecteur Générique [6] : détecteur similaire à celui proposé par Dalal *et al* qui a été entrainé sur la base INRIA.
- Détecteur Cible avec étiquetage manuel : c'est un détecteur entrainé sur un ensemble d'échantillons étiqueté manuellement et composé de tout les piétons présents dans les images de spécialisation et des imagettes négatives extraites de manière aléatoire.
- Wang 2012 [5] : Il s'agit d'un détecteur spécifique à la scène cible, entrainé à la fois sur des échantillons de la base INRIA et de la scène cible. Les échantillons cibles et sources sélectionnées sont ceux qui ont des scores de



(a)

(b)

FIGURE 4 – Performance Globale : (a) Comparaison avec d'autres méthodes d'état de l'art, (b) Effet de spécialisation : (gauche) résultat de détecteur générique et (droite) résultat de détecteur spécialisé

confiance élevés. Les scores sont calculés à l'aide de plusieurs indices contextuels et le choix des échantillons d'apprentissage est assuré par une nouvelle variante des SVM dite "Confidence-Encoded SVM" qui favorise les échantillons avec scores élevés.

– Nair 2004 [9] : Un détecteur issu d'une approche d'adaptation automatique qui sélectionne des échantillons cibles à ajouter dans la base d'apprentissage initiale à l'aide d'un algorithme d'extraction fond-forme (détecteur similaire à celui proposé dans [9] mais crée avec le descripteur HOG et le classifieur SVM)

Dans la suite, l'indication d'un taux de détection est toujours relié à un seul faux positif par image (FPPI). Les figures 4.(a) et 4.(b) montrent qu'un détecteur spécialisé dépasse significativement un détecteur générique. La performance sur l'ensemble d'images de test de la base CUHK_Square a été améliorée de 26.6% à 74.37%. Le détecteur spécialisé dépasse également les deux autres détecteurs de Nair 2004 et Wang 2012 respectivement de 45.57% et 23,25%. Par contre, le détecteur cible avec étiquetage manuel dépasse légèrement le détecteur spécialisé pour un FPPI inférieur à 0.2 mais le détecteur spécialisé dépasse nettement ce dernier pour un FPPI strictement supérieur à 0.2. En particulier, il présente un taux d'amélioration égale à 23,25% pour FPPI=1.

L'illustration de la FIGURE 4.(b) montre clairement que la spécialisation a réduit le nombre de fausses détections de manière considérable et qu'elle a amélioré la détection de certains piétons non détectés par le détecteur générique.

5 Conclusion et perspectives

Dans ce papier, nous avons proposé une nouvelle méthode de transfert d'apprentissage transductif basée sur un filtre séquentiel de Monte Carlo qui a permis de spécialiser un classifieur générique à une scène de trafic urbain sans étiquetage manuel de données. La méthode a donné de bons résultats sur des données réelles non seulement par rapport au classifieur générique mais par rapport à d'autres méthodes de transfert telles que celle de Wang (qui utilise une stratégie d'observation beaucoup plus sophistiquée que la nôtre). De plus, elle converge dès les premières itérations (trois ou quatre itérations sont suffisantes).
Nous avons utilisé un détecteur à base des HOG et SVM mais notre méthode est générique et peut être étendue à d'autres types de détecteurs. Nous envisageons d'améliorer notre fonction d'observation pour une meilleure sélection des échantillons d'apprentissage, de valider sur d'autres bases, d'optimiser le temps d'exécution au cours d'une itération et d'étendre la spécialisation à des problèmes multi-classes.

Remerciements

Ce travail est financé dans le cadre d'une convention Cifre avec la société Logiroad et a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'avenir dans le cadre du projet LabEx IMobS3 (ANR-10-LABX-16-01), d'une aide de l'Union Européenne au titre du Programme Compétitivité Régionale et Emploi 2007-2013 (FEDER - Région Auvergne) et d'une aide de la Région Auvergne.

Références

[1] A. Doucet, N. de Feitas et N. Gordon. Sequential Monte Carlo methods in practice. Springer, 2001.

[2] A. Levin, P. Viola et Y. Freund. Unsupervised improvement of visual detectors using cotraining. In Proc. CV, pp 626-633, 2003.

[3] C. Rosenberg, M. Hebert et H. Schneiderman : Semi-supervised self-training of object detection models, In IEEE workshop on ACV, 2005.

[4] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, et A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge, IJCV, 88(2), pp 303-338, 2010.

[5] M. Wang, W. Li et X. Wang. Transferring a generic pedestrian detector towards specific scenes, In Proc. CVPR, 2012.

[6] N. Dalal et B. Triggs. Histograms of oriented gradients for human detection. In Proc. CVPR, 1, pp 886-893, 2005.

[7] S. J. Pan et Q. Yang. A survey on transfer learning. In Trans. IEEE KDE, 22(10), pp 1345-1359, 2010.

[8] T. Chesnais T., N. Allezard, Y. Dhome et T. Chateau. Automatic process to build a contextualized detector. In VISAPP, 1, pp 513-520, 2012.

[9] V. Nair et J. J. Clark. An unsupervised, online learning framework for moving object detection. In Proc. CVPR, 2, pp 317-324, 2004.