

# Apprentissage de dictionnaire pour un modèle d'apparence parcimonieux en suivi visuel

Sylvain ROUSSEAU<sup>1</sup>, Christelle GARNIER<sup>2</sup>, Pierre CHAINAIS<sup>1</sup>

<sup>1</sup>École Centrale de Lille  
CRIStAL, UMR CNRS 9189, Villeneuve d'Ascq, France

<sup>2</sup>Institut Mines-Telecom / Telecom Lille  
CRIStAL, UMR CNRS 9189, Villeneuve d'Ascq, France

sylvain.rousseau@ec-lille.fr, christelle.garnier@telecom-lille.fr, pierre.chainais@ec-lille.fr

**Résumé** – Cet article présente une nouvelle approche pour le suivi visuel par filtrage particulière. L'apparence de l'objet cible est décrite par une représentation parcimonieuse fournie par apprentissage de dictionnaire, ce qui permet de créer un modèle de dimension réduite. La vraisemblance d'une région candidate est construite à partir d'une mesure de similarité qui s'interprète comme le résultat d'un filtrage adapté dans le nouvel espace de représentation formé par le dictionnaire. Cette approche permet de détecter de manière optimale la présence de l'ensemble des patchs de référence extraits de la cible aux positions considérées dans la région candidate. La validation expérimentale montre l'efficacité et la robustesse de l'approche proposée.

**Abstract** – This paper presents a novel approach to visual object tracking based on particle filtering. The appearance of the target object is described by a sparse representation provided by dictionary learning, which leads to a model of reduced dimension. The likelihood of a candidate region is built on a similarity measure which can be interpreted as the result of a matched filter in the new representation space formed by the dictionary. Thus it can optimally detect a set of reference patches extracted from the target at known positions in the candidate region. Experimental validation shows the efficiency and the robustness of the proposed approach.

## 1 Introduction

Le suivi d'objets dans une séquence vidéo est une tâche essentielle en vision par ordinateur [7]. Parmi les approches proposées, le filtrage particulière [2] a rencontré un vif succès. Il s'agit surtout de construire un modèle d'apparence capable de caractériser la cible de manière discriminante et robuste au fil du temps. Avec l'essor des représentations parcimonieuses, de nouveaux modèles sont apparus. Un modèle d'apparence global s'inspirant de l'approche développée par [6] pour la reconnaissance faciale a d'abord été proposé dans [4]. Le dictionnaire est constitué de deux parties : des images de référence représentant la cible et des modèles triviaux pour prendre en compte les pixels bruités. Chaque région candidate est décomposée de façon parcimonieuse dans le dictionnaire complet et sa vraisemblance est reliée à la qualité de sa représentation dans le dictionnaire des images de référence.

Pour mieux prendre en compte les occultations partielles, un modèle d'apparence local fondé sur l'extraction de patchs et leur décomposition dans un dictionnaire est plus adapté. Des descripteurs sont construits à partir des codes parcimonieux des patchs et la vraisemblance d'une région candidate repose sur une comparaison entre les

descripteurs qui la caractérisent et ceux de l'image de référence. Plusieurs descripteurs ont été envisagés. Le descripteur moyen [8] est construit par moyennage du code parcimonieux de chaque patch. Le descripteur maximum [5] consiste à sélectionner la valeur maximale du code parcimonieux de chaque patch. Les deux méthodes sont relativement robustes mais présentent l'inconvénient de ne pas tenir compte du lieu des patchs lors de la construction du descripteur, ce qui les rend très imprécises. Le descripteur aligné, développé dans [3], conserve une information sur la structure spatiale de la cible en sélectionnant l'élément du code parcimonieux qui correspond à l'endroit où le patch a été extrait. Cependant le dictionnaire utilisé consiste simplement en la juxtaposition de patchs extraits d'une ou de plusieurs images de référence suivant une grille spatiale fixée a priori, ce qui limite la robustesse de cette approche.

Dans cet article, nous proposons de combiner le principe du descripteur aligné avec un apprentissage de dictionnaire pour représenter l'apparence de l'objet à suivre avec un modèle de dimension réduite. La log-vraisemblance d'une région candidate est construite par comparaison des codes parcimonieux qui s'interprète comme un filtrage adapté. La validation expérimentale sur quelques séquences vidéo montre l'efficacité et la robustesse de l'approche proposée.

## 2 Suivi par descripteur aligné

Nous rappelons d'abord la construction du descripteur aligné [3]. La cible est potentiellement décrite par plusieurs images de référence. Pour simplifier, nous n'en considérons qu'une seule, notée  $T$ . Un ensemble de  $N$  patches de taille  $c^2$  est extrait de  $T$  à des positions prédéfinies suivant une structure spatiale donnée (grille de pas constant). Les  $N$  patches  $(p_i)_{i=1}^N$  sont transformés en vecteurs et concaténés dans une matrice  $\mathbf{L}_p = [p_1, \dots, p_N] \in \mathbb{R}^{c^2 \times N}$ , qui est utilisée comme dictionnaire par la suite.

Pour caractériser une région candidate  $C$ , on commence par la redimensionner pour qu'elle soit de même taille que l'image de référence  $T$ . Ensuite, on extrait un nombre  $N$  de patches  $(q_i)_{i=1}^N$  suivant la même structure spatiale que précédemment. Le code parcimonieux de chacun des patches  $(q_i)_{i=1}^N$  est alors calculé dans le dictionnaire  $\mathbf{L}_p$  de taille  $N$  par la minimisation suivante,

$$\arg \min_{u_i} \|q_i - \mathbf{L}_p u_i\|_2^2 \quad \text{tel que} \quad \|u_i\|_1 \leq \rho, \quad (1)$$

où le paramètre  $\rho$  contrôle la parcimonie du code  $u_i \in \mathbb{R}^N$ . Ainsi, chaque patch  $q_i$  est approché par une combinaison parcimonieuse des patches  $(p_i)_{i=1}^N$ . Le descripteur aligné de  $u_i$  consiste à sélectionner la contribution du patch  $p_i$  situé au même endroit que le patch  $q_i$ , ce qui correspond au  $i$ -ème élément  $u_{ii}$  de  $u_i$ . En notant  $\mathbf{U} \in \mathbb{R}^{N \times N}$  la matrice carrée qui regroupe les  $u_i$ , le descripteur aligné est la diagonale de  $\mathbf{U}$ . Finalement, la log-vraisemblance  $\mathcal{L}(C, T)$  d'une région candidate  $C$  par rapport à la référence  $T$  est définie par,

$$\mathcal{L}(C, T) = \text{Tr} \mathbf{U} = \sum_{i=1}^N u_{ii}. \quad (2)$$

La modélisation de l'apparence d'une cible par le descripteur aligné est plus précise que par les descripteurs moyen et maximum car elle prend en compte le lieu où les patches ont été extraits. L'inconvénient majeur réside dans le manque de flexibilité de la méthode. La taille du dictionnaire est nécessairement égale au nombre de patches extraits. Pour extraire plus de patches ou changer les lieux d'extraction pour gagner en précision, il faut reconstruire le dictionnaire et augmenter sa taille. Le tracker s'avère de plus modérément robuste.

## 3 Approche proposée

Pour surmonter les limitations du descripteur aligné, nous proposons d'apprendre un dictionnaire  $\mathbf{D}_p \in \mathbb{R}^{c^2 \times K}$  de taille  $K \ll N$  à partir des patches  $\mathbf{L}_p = [p_1, \dots, p_N] \in \mathbb{R}^{c^2 \times N}$  extraits de l'image de référence  $T$ . L'apprentissage est effectué par la minimisation suivante,

$$\arg \min_{\mathbf{D}_p, \mathbf{Z}} \|\mathbf{L}_p - \mathbf{D}_p \mathbf{Z}\|_F^2 \quad \text{tel que} \quad \|z_i\|_1 \leq \rho, \quad \forall i, \quad (3)$$

où  $z_i$  est la  $i$ -ème colonne de la matrice  $\mathbf{Z} \in \mathbb{R}^{K \times N}$  et  $\|\cdot\|_F$  est la norme de Frobenius. Ainsi,  $z_i \in \mathbb{R}^K$  est le code

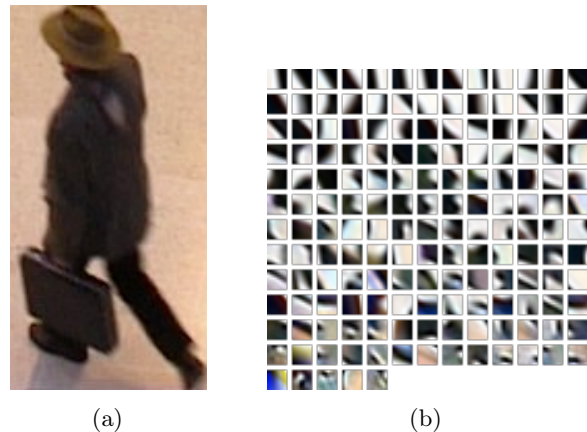


FIGURE 1 – (a) Image de référence  $T$ . (b) Atomes du dictionnaire  $\mathbf{D}_p$  ordonnés par contribution décroissante.

parcimonieux du patch  $p_i$  dans le dictionnaire  $\mathbf{D}_p$ . Grâce à la réduction de dimension apportée par l'utilisation d'un dictionnaire, le nombre  $N$  de patches extraits peut être choisi grand. Par la suite, l'apprentissage est effectué une seule fois avant le suivi sur tous les patches disponibles dans l'image de référence. La Figure 1 illustre un dictionnaire de taille 161 appris sur les 16109 patches de taille  $8 \times 8$  extraits d'une image de référence de taille  $188 \times 96$ .

Pour décrire une région candidate,  $n$  patches  $(q_i)_{i \in J}$  sont extraits à des positions décrites par l'ensemble  $J : J \subset \{1, \dots, N\}$  et  $\#J = n$ . Contrairement à la méthode précédente, le choix du nombre  $n$  de patches et de leurs positions décrites par  $J$  est complètement libre et peut même varier au cours du temps. Les codes parcimonieux  $\mathbf{V} = [v_i]_{i \in J} \in \mathbb{R}^{K \times n}$  de ces patches dans le dictionnaire  $\mathbf{D}_p$  sont calculés par la minimisation suivante,

$$\arg \min_{v_i} \|q_i - \mathbf{D}_p v_i\|_2^2 \quad \text{tel que} \quad \|v_i\|_1 \leq \rho. \quad (4)$$

Pour construire la log-vraisemblance d'une région candidate  $C$ , on compare les représentations parcimonieuses dans le dictionnaire  $\mathbf{D}_p$  des  $n$  patches extraits de  $C$  et  $T$  aux mêmes emplacements, définis par  $J$ . Pour cela, on effectue un filtrage adapté qui optimise le rapport signal sur bruit en détection. Ainsi, la log-vraisemblance d'une région candidate  $C$  caractérisée par son codage parcimonieux  $\mathbf{V}$  par rapport à l'image de référence  $T$  de codage parcimonieux  $\mathbf{Z}_J = [z_i]_{i \in J}$  est définie comme une intercorrélacion,

$$\frac{1}{\|\mathbf{Z}_J\|_F} \text{Tr} \mathbf{Z}_J^T \mathbf{V}. \quad (5)$$

Cette log-vraisemblance détecte de manière optimale la présence des  $n$  patches de référence représentés par  $\mathbf{Z}_J$  dans la région candidate aux positions considérées.

Cette formulation généralise la modélisation d'apparence par le descripteur aligné proposé dans [3]. En l'absence d'apprentissage de dictionnaire, c'est-à-dire dans le cas particulier où le dictionnaire résulte de la concaténation des patches de référence :  $\mathbf{D}_p = \mathbf{L}_p$ , le codage de ces

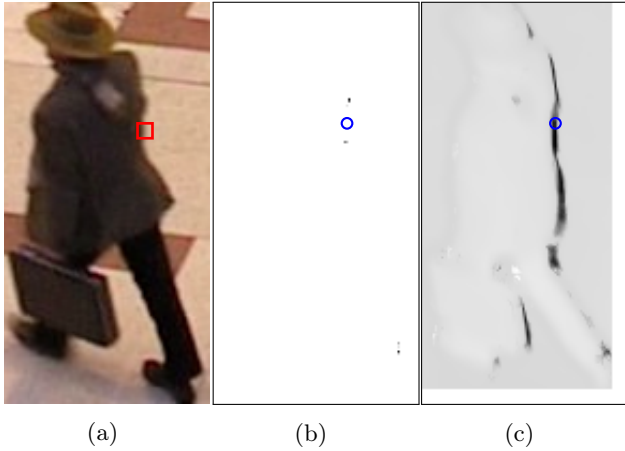


FIGURE 2 – (a) Patch  $q_i$  extrait d’une région candidate. (b) Codage parcimonieux  $u_i$  du patch. (c) Intercorrélation  $\mathbf{Z}_J^T v_i$ .

patches devient trivial :  $\mathbf{Z}_J = \mathbf{I}$  et l’expression (5) devient identique à l’expression (2).

Il reste à prendre en compte les valeurs moyennes  $\bar{p}_i$  et  $\bar{q}_i$  des patches  $p_i$  et  $q_i$  extraits de  $T$  et  $C$  qui sont ignorées lors de l’apprentissage du dictionnaire et du codage parcimonieux (qui ne décrivent que les variations). La log-vraisemblance (5) devient finalement

$$\mathcal{L}_J(C, T) = \frac{1}{\|\mathbf{Z}_J\|_F} \text{Tr} \mathbf{Z}_J^T \mathbf{V} - \lambda \sum_{i \in J} |\bar{p}_i - \bar{q}_i|^2, \quad (6)$$

où  $\lambda > 0$  est un coefficient de pondération ( $\lambda = 1$  par défaut).

La Figure 2 permet de comparer les deux approches. Un patch  $q_i$  est extrait d’une région candidate redimensionnée à la Figure (2a). Les figures suivantes représentent sous forme d’images la similarité entre ce patch  $q_i$  et les patches  $(p_i)_{i \in J}$  avec  $J = \{1, \dots, N\}$  extraits de l’image de référence. Sans apprentissage de dictionnaire à la Figure (2b), la similarité correspond directement au code parcimonieux  $u_i$  du patch  $q_i$  dans le dictionnaire trivial  $\mathbf{L}_p$ . Avec apprentissage de dictionnaire à la Figure (2c), elle combine l’intercorrélacion  $\mathbf{Z}_J^T v_i$  où  $v_i$  est le code parcimonieux du patch  $q_i$  dans le dictionnaire  $\mathbf{D}_p$  et la différence d’intensité moyenne via  $|\bar{p}_i - \bar{q}_i|^2$ . Le descripteur aligné ne prend en compte que la similarité avec le patch de référence  $p_i$  situé à la même position que  $q_i$ . La valeur extraite est indiquée par le cercle bleu. Avec le descripteur aligné classique, la similarité estimée entre les patches  $q_i$  et  $p_i$  est très faible. Un grand nombre d’atomes (de patches) de  $\mathbf{L}_p$  sont similaires et le code parcimonieux  $u_i$  du patch  $q_i$  n’inclut pas nécessairement l’atome  $p_i$ . L’approche proposée fournit de meilleurs résultats puisqu’une similarité est détectée dès que les décompositions parcimonieuses des patches  $q_i$  et  $p_i$  partagent quelques atomes du dictionnaire  $\mathbf{D}_p$ , ce qui est facteur de robustesse.

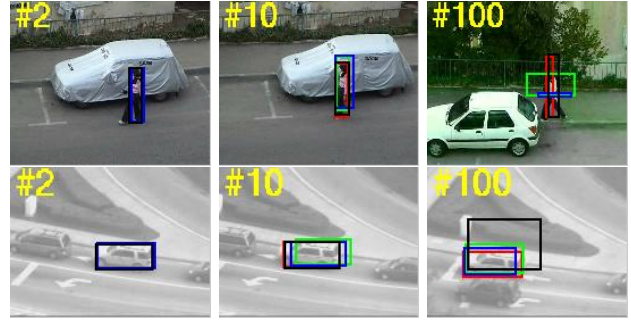


FIGURE 3 – Résultats qualitatifs de suivi sur les séquences Woman et Suv pour différents descripteurs : proposé en rouge, maximal en vert, moyen en bleu et aligné en noir.

## 4 Algorithme de suivi

Le suivi de l’objet est effectué par filtrage particulière [2]. Soit une séquence d’observations  $y_{1:k} = (y_1, \dots, y_k)$  jusqu’à l’instant  $k$ . L’état  $x_k$  est estimé à l’aide de la densité a posteriori  $p(x_k | y_{1:k})$  approchée par un ensemble de  $N_p$  particules pondérées,  $\{x_k^{(m)}, w_k^{(m)}\}_{m=1}^{N_p}$ . Le vecteur d’état s’écrit  $x_k = [c_k, d_k]^T$  où  $c_k = [c_k^x, c_k^y]$  est la position du coin supérieur gauche et  $d_k = [d_k^x, d_k^y]$  la taille de la fenêtre de suivi. Les particules  $x_k^{(m)}$  sont propagées selon le modèle dynamique, ici une marche aléatoire Gaussienne :  $x_k | x_{k-1} \sim \mathcal{N}(x_{k-1}, \Sigma)$  où  $\Sigma$  est la matrice de covariance diagonale qui définit l’espace de recherche autour de l’état précédent. Les poids  $w_k^{(m)}$  sont mis à jour selon l’expression récursive,  $w_k^{(m)} = w_{k-1}^{(m)} \cdot p(y_k | x_k^{(m)})$  où la vraisemblance  $p(y_k | x_k)$  mesure l’adéquation des observations à l’état. La vraisemblance est définie directement à partir de l’expression (6),

$$p(y_k | x_k) \propto \exp \{ \mu \cdot \mathcal{L}_J(C, T) \}, \quad (7)$$

où  $C$  est la région candidate identifiée par  $x_k$  et  $\mu$  un paramètre fixé expérimentalement. Finalement, après normalisation des poids et, si nécessaire, rééchantillonnage des particules, l’état final est obtenu par l’estimateur suivant,

$$\hat{x}_k = \mathbb{E}[x_k | y_{1:k}] = \sum_{m=1}^{N_p} w_k^{(m)} x_k^{(m)}. \quad (8)$$

## 5 Validation expérimentale

Quatre descripteurs sont comparés : maximal, moyen, aligné ainsi que l’approche proposée. Les problèmes d’optimisation (1), (3) et (4) sont résolus en utilisant la bibliothèque Spams<sup>1</sup> avec  $\rho = 0.25$ . Les patches considérés sont distribués selon une grille régulière de pas 8. Pour le filtrage particulière,  $N_p = 600$  particules,  $\mu = 4.6$  et  $\Sigma = \text{diag}(20, 20, 4, 4)$ . La taille du dictionnaire est fixée à  $K = 64$ .

1. <http://spams-devel.gforge.inria.fr/>

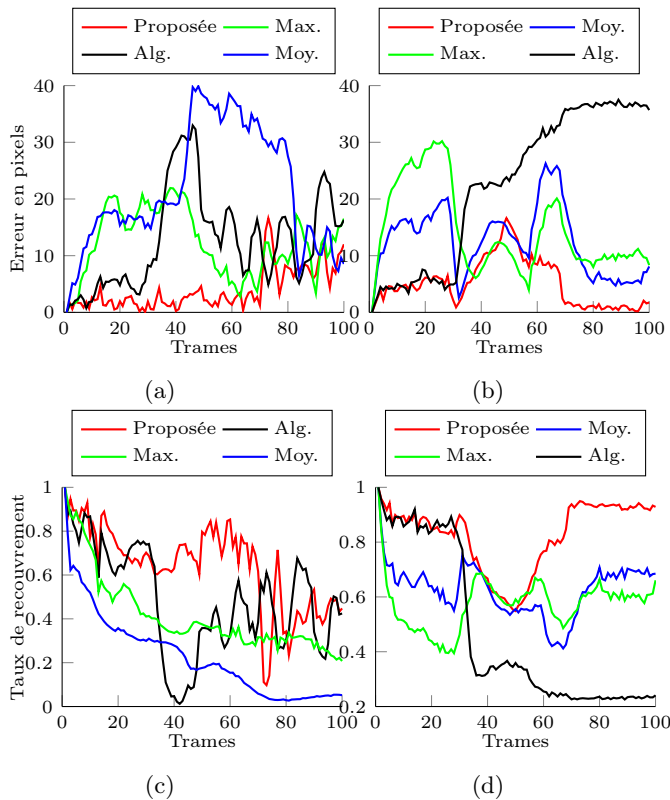


FIGURE 4 – Résultats quantitatifs sur les séquences vidéo *Woman* et *Suv*. Erreur de localisation sur : (a) *Woman*, (b) *Suv*. Taux de recouvrement sur : (c) *Woman*, (d) *Suv*.

Les descripteurs sont testés sur les 100 premières images des séquences *Suv* et *Woman*. La Figure 3 montre quelques résultats qualitatifs sur les 2<sup>ème</sup>, 10<sup>ème</sup> et 100<sup>ème</sup> images. La Figure 4 montre les résultats quantitatifs obtenus. La première ligne représente l’erreur de localisation pour chaque image. Il s’agit de la distance euclidienne entre le centre de la fenêtre estimée et le centre de la fenêtre correspondant à la vérité terrain. La deuxième ligne représente le taux de recouvrement [1] qui est le rapport entre l’aire de l’intersection et de l’union des deux fenêtres.

Les descripteurs maximal et moyen donnent les plus mauvais résultats de suivi dès les premières images des deux séquences car leur vraisemblance est construite sans tenir compte des lieux des patches extraits. En revanche, ils montrent une certaine robustesse sur la séquence *Suv* en retrouvant la cible après une sortie partielle de l’image. À l’inverse, le descripteur aligné est plus précis car il conserve une information spatiale sur la cible lors de la construction de la vraisemblance. Mais la qualité du suivi se dégrade rapidement car cette approche compare directement les patches extraits des régions candidates aux patches de l’image de référence. Comme le montre la Figure (2b), il est possible de ne pas détecter de similarité entre les patches alignés. La méthode proposée permet d’obtenir les meilleures performances de suivi. Elle est aussi précise que le descripteur aligné parce qu’elle prend en compte les positions des

patches ; elle est aussi plus robuste grâce à l’apprentissage de dictionnaire car elle détecte une similarité entre deux patches dès qu’ils partagent un atome du dictionnaire dans leur décomposition parcimonieuse.

## 6 Conclusion

Nous avons proposé un modèle de dimension réduite basé sur un codage parcimonieux de patches dans un dictionnaire appris pour caractériser l’apparence d’un objet dans une séquence d’images. Pour le suivi par filtrage particulière, la vraisemblance est calculée par filtrage adapté, ce qui permet de détecter de manière optimale la présence des patches de référence extraits de l’objet dans une région candidate à des positions choisies. Les expériences numériques montrent la précision de notre approche liée au principe d’alignement des patches et sa robustesse grâce à l’apprentissage de dictionnaire.

## Références

- [1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2) :303–338, 2010.
- [2] M. Isard and A. Blake. Condensation–conditional density propagation for visual tracking. *International journal of computer vision*, 29(1) :5–28, 1998.
- [3] X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1822–1829. IEEE, 2012.
- [4] X. Mei and H. Ling. Robust visual tracking using  $\ell_1$  minimization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1436–1443. IEEE, 2009.
- [5] Q. Wang, F. Chen, J. Yang, W. Xu, and M.-H. Yang. Transferring visual prior for online object tracking. *IEEE Transactions on Image Processing*, 21(7) :3296–3305, 2012.
- [6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(2) :210–227, 2009.
- [7] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song. Recent advances and trends in visual tracking : A review. *Neurocomputing*, 74(18) :3823–3831, 2011.
- [8] S. Zhang, H. Yao, and S. Liu. Robust visual tracking using feature-based visual attention. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1150–1153. IEEE, 2010.