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

Sylvain Rousseau[1], Christelle Garnier[2], Pierre Chainais[1]

[1]École Centrale de Lille
CRIStAL, UMR CNRS 9189, Villeneuve d'Ascq, France

[2]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 particulaire. 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 [1]. Parmi les approches proposées, le filtrage particulaire [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 [3] 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 [5] est construit par moyennage du code parcimonieux de chaque patch. Le descripteur maximum [6] 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 [7], 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 [7]. 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 patchs de taille c^2 est extrait de T à des positions prédéfinies suivant une structure spatiale donnée (grille de pas constant). Les N patchs $(p_i)_{i=1}^N$ sont transformés en vecteurs et concaténés dans une matrice $\boldsymbol{L_p} = [p_1, \ldots, p_N] \in \mathbb{R}^{c^2 \times N}$, 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 patchs $(q_i)_{i=1}^N$ suivant la même structure spatiale que précédemment. Le code parcimonieux de chacun des patchs $(q_i)_{i=1}^N$ est alors calculé dans le dictionnaire $\boldsymbol{L_p}$ de taille N par la minimisation suivante,

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

où le paramètre ρ contrôle la parcimonie du code $u_i \in \mathbb{R}^N$. Ainsi, chaque patch q_i est approché par une combinaison parcimonieuse des patchs $(p_i)_{i=1}^N$. 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 $\boldsymbol{U} \in \mathbb{R}^{N \times N}$ la matrice carrée qui regroupe les u_i, le descripteur aligné est la diagonale de $\boldsymbol{U}$. Finalement, la log-vraisemblance $\mathcal{L}(C, T)$ d'une région candidate C par rapport à la référence T est définie par,

$$\mathcal{L}(C, T) = \operatorname{Tr} \boldsymbol{U} = \sum_{i=1}^{N} u_{ii}. \qquad (2)$$

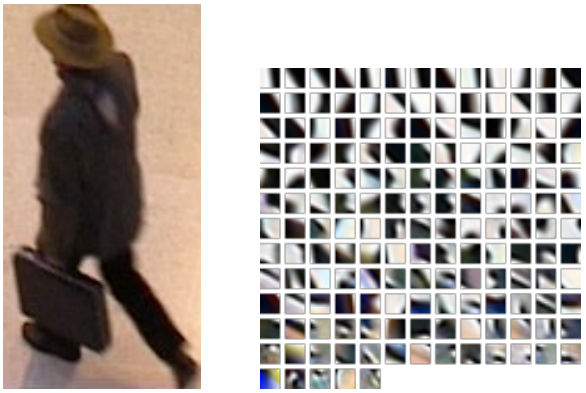
La modélisation de l'apparence d'une cible par le descripteur aligné est plus précise que par les descripteurs moyen et maximum car elle prend en compte le lieu où les patchs ont été extraits. L'inconvénient majeur réside dans le manque de flexibilité de la méthode. La taille du dictionnaire est nécessairement égale au nombre de patchs extraits. Pour extraire plus de patchs 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 $\boldsymbol{D_p} \in \mathbb{R}^{c^2 \times K}$ de taille $K \ll N$ à partir des patchs $\boldsymbol{L_p} = [p_1, \ldots, p_N] \in \mathbb{R}^{c^2 \times N}$ extraits de l'image de référence T. L'apprentissage est effectué par la minimisation suivante,

$$\arg\min_{\boldsymbol{D_p}, \boldsymbol{Z}} \|\boldsymbol{L_p} - \boldsymbol{D_p} \boldsymbol{Z}\|_F^2 \quad \text{tel que} \quad \|z_i\|_1 \leq \rho, \ \forall i, \qquad (3)$$

où z_i est la i-ème colonne de la matrice $\boldsymbol{Z} \in \mathbb{R}^{K \times N}$ et $\|\cdot\|_F$ est la norme de Frobenius. Ainsi, $z_i \in \mathbb{R}^K$ est le code parcimonieux du patch p_i dans le dictionnaire $\boldsymbol{D_p}$. Grâce à la réduction de dimension apportée par l'utilisation d'un dictionnaire, le nombre N de patchs extraits peut être choisi grand. Par la suite, l'apprentissage est effectué une seule fois avant le suivi sur tous les patchs disponibles dans l'image de référence. La Figure 1 illustre un dictionnaire de taille 161 appris sur les 16109 patchs de taille 8×8 extraits d'une image de référence de taille 188×96.



(a) (b)

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

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

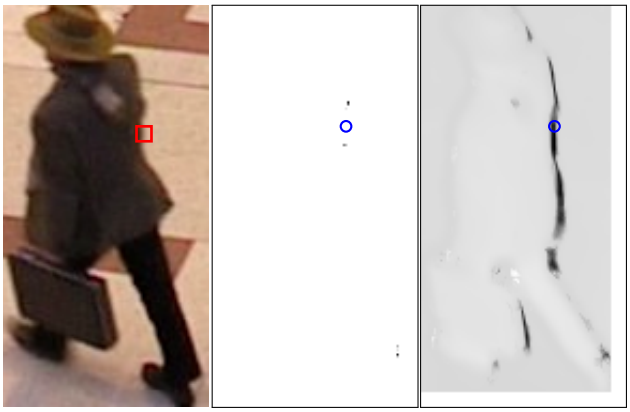
$$\arg\min_{v_i} \|q_i - \boldsymbol{D_p} v_i\|_2^2 \quad \text{tel que} \quad \|v_i\|_1 \leq \rho. \qquad (4)$$

Pour construire la log-vraisemblance d'une région candidate C, on compare les représentations parcimonieuses dans le dictionnaire $\boldsymbol{D_p}$ des n patchs 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 $\boldsymbol{V}$ par rapport à l'image de référence T de codage parcimonieux $\boldsymbol{Z_J} = [z_i]_{i \in J}$ est définie comme une intercorrélation,

$$\frac{1}{\|\boldsymbol{Z_J}\|_F} \operatorname{Tr} \boldsymbol{Z_J^T} \boldsymbol{V}. \qquad (5)$$

Cette log-vraisemblance détecte de manière optimale la présence des n patchs de référence représentés par $\boldsymbol{Z_J}$ dans la région candidate aux positions considérées.

Cette formulation généralise la modélisation d'apparence par le descripteur aligné proposé dans [7]. En l'absence d'apprentissage de dictionnaire, c'est-à-dire dans le cas particulier où le dictionnaire résulte de la concaténation des patchs de référence : $\boldsymbol{D_p} = \boldsymbol{L_p}$, le codage de ces



(a) (b) (c)

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

patchs devient trivial : $\boldsymbol{Z}_{\boldsymbol{J}} = \boldsymbol{I}$ 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 patchs 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

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

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 patchs $(p_i)_{i \in J}$ 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 $\boldsymbol{L}_{\boldsymbol{p}}$. Avec apprentissage de dictionnaire à la Figure (2c), elle combine l'intercorrélation $\boldsymbol{Z}_{\boldsymbol{J}}^T v_i$ où v_i est le code parcimonieux du patch q_i dans le dictionnaire $\boldsymbol{D}_{\boldsymbol{p}}$ 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 patchs q_i et p_i est très faible. Un grand nombre d'atomes (de patchs) de $\boldsymbol{L}_{\boldsymbol{p}}$ 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 patchs q_i et p_i partagent quelques atomes du dictionnaire $\boldsymbol{D}_{\boldsymbol{p}}$, ce qui est facteur de robustesse.

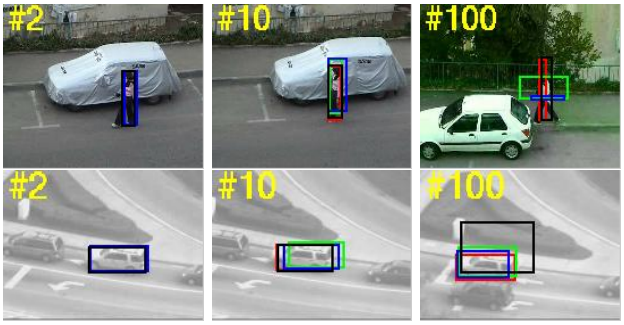


FIGURE 3 – Résultats qualitatifs de suivi sur les séquences `Woman` et `Suv` pour différents descripteurs : proposé en rouge, maximal en vert, moyen en bleu et aligné en noir.

4 Algorithme de suivi

Le suivi de l'objet est effectué par filtrage particulaire [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 Σ 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),

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

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

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

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 [1] avec $\rho = 0.25$. Les patchs considérés sont distribués selon une grille régulière de pas 8. Pour le filtrage particulaire, $N_p = 600$ particules, $\mu = 4.6$ et $\Sigma = \operatorname{diag}(20, 20, 4, 4)$. La taille du dictionnaire est fixée à $K = 64$.

1. `http://spams-devel.gforge.inria.fr/`

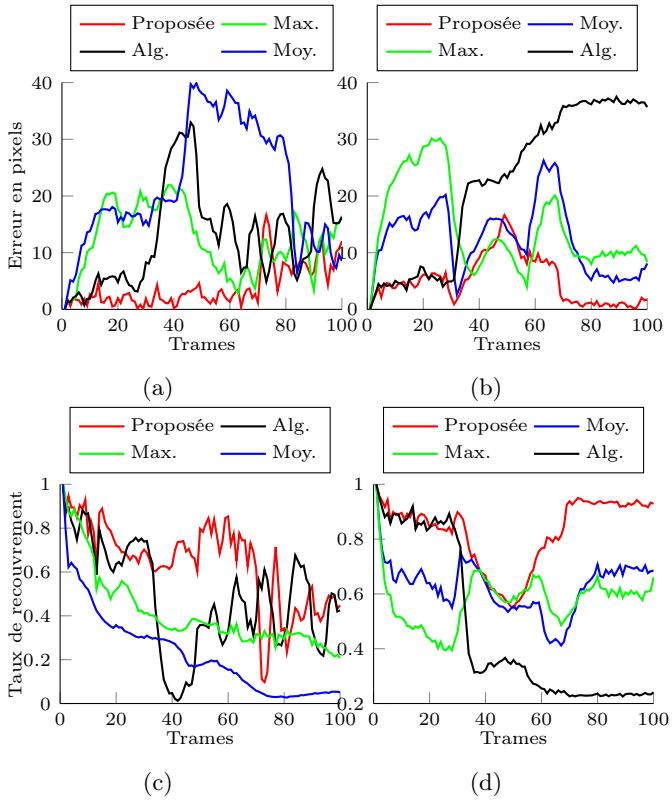


FIGURE 4 – Résultats quantitatifs sur les séquences vidéo `Woman` et `Suv`. Erreur de localisation sur : (a) `Woman`, (b) `Suv`. Taux de recouvrement sur : (c) `Woman`, (d) `Suv`.

Les descripteurs sont testés sur les 100 premières images des séquences `Suv` et `Woman`. La Figure 3 montre quelques résultats qualitatifs sur les $2^{\text{ème}}$, $10^{\text{ème}}$ et $100^{\text{ème}}$ 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 [8] 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 patchs 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 patchs extraits des régions candidates aux patchs de l'image de référence. Comme le montre la Figure (2b), il est possible de ne pas détecter de similarité entre les patchs 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 patchs ; elle est aussi plus robuste grâce à l'apprentissage de dictionnaire car elle détecte une similarité entre deux patchs 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 patchs dans un dictionnaire appris pour caractériser l'apparence d'un objet dans une séquence d'images. Pour le suivi par filtrage particulaire, la vraisemblance est calculée par filtrage adapté, ce qui permet de détecter de manière optimale la présence des patchs 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 patchs et sa robustesse grâce à l'apprentissage de dictionnaire.

Références

[1] H. YANG, L. SHAO, F. ZHENG, L. WANG et Z. SONG, "Recent advances and trends in visual tracking : a review", *Neurocomputing*, t. 74, n° 18, p. 3823–3831, 2011.

[2] M. ISARD et A. BLAKE, "Condensation–conditional density propagation for visual tracking", *International journal of computer vision*, t. 29, n° 1, p. 5–28, 1998.

[3] J. WRIGHT, A. Y. YANG, A. GANESH, S. S. SASTRY et Y. MA, "Robust face recognition via sparse representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, t. 31, n° 2, p. 210–227, 2009.

[4] X. MEI et H. LING, "Robust visual tracking using ℓ_1 minimization", in *IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2009, p. 1436–1443.

[5] S. ZHANG, H. YAO et S. LIU, "Robust visual tracking using feature-based visual attention", in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, IEEE, 2010, p. 1150–1153.

[6] Q. WANG, F. CHEN, J. YANG, W. XU et M.-H. YANG, "Transferring visual prior for online object tracking", *IEEE Transactions on Image Processing*, t. 21, n° 7, p. 3296–3305, 2012.

[7] X. JIA, H. LU et M.-H. YANG, "Visual tracking via adaptive structural local sparse appearance model", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012, p. 1822–1829.

[8] M. EVERINGHAM, L. VAN GOOL, C. K. WILLIAMS, J. WINN et A. ZISSERMAN, "The pascal visual object classes (voc) challenge", *International journal of computer vision*, t. 88, n° 2, p. 303–338, 2010.