Connexions entre descripteurs locaux et globaux pour la reconnaissance d'objets dans les vidéos

Bruno Lameyre Valérie Gouet-Brunet
CEDRIC/CNAM - 292, rue Saint-Martin - F75141 Paris Cedex 03
{bruno.lameyre,valerie.gouet}@cnam.fr

Concours Jeune Chercheur : Oui

Résumé

Dans ce travail, nous présentons une approche de reconnaissance d'objets génériques à partir de flux vidéos, basée sur la construction d'un catalogue de caractéristiques visuelles hétérogènes. Notre première contribution porte sur la description de l'apparence visuelle des objets, en proposant l'utilisation conjointe de primitives génériques et complémentaires de différentes natures : d'un côté, un ensemble de descripteurs locaux, aux propriétés bien connues, telle leur robustesse à l'arrière-plan ; de l'autre côté, un contour actif comme descripteur global, fournissant une description haut-niveau de la forme de l'objet. Notre seconde contribution propose de structurer efficacement ces descripteurs, notamment en établissant des connexions entre eux. Cette approche est comparée à une approche classique et évaluée sur plusieurs séquences contenant 20 objets. Nous montrons sa pertinence pour l'annotation automatique de contenus vidéo, où de bons taux de reconnaissance sont atteints, tout en préservant des performances compatibles avec le temps-réel.

Mots clefs

Indexation d'images par contenu visuel, Reconnaissance d'objets, Descripteurs locaux, Contours actifs, Flux vidéo.

1 Introduction

La reconnaissance d'objets est depuis longtemps un domaine de recherche actif. La plupart des approches rencontrées base la phase d'apprentissage sur un ensemble d'images fixes. Depuis peu, un petit nombre d'approches propose d'exploiter la richesse de la vidéo, sur la base des observations suivantes : les humains reconnaissent mieux un objet quand il est en mouvement plutôt qu'à partir de simples vues ; techniquement, une vidéo fournit de multiples vues de l'objet, facilement reliables par une méthode de suivi entre trames consécutives. Ces approches exploitent l'information temporelle des vidéos en extrayant une ou plusieurs primitives visuelles dans chaque trame et en le(s) suivant le long de la séquence. Suivre une primitive permet notamment de produire une description plus robuste en modélisant sa variabilité le long de la trajectoire et en structurant l'espace de description engendré pour regrouper les caractéristiques redondantes par objet et entre objets. Dans [1] par exemple, des modèles 3D sont construits en exploitant le suivi de patches à partir d'objets en mouvement dans les vidéos. Dans [2], une approche probabiliste de suivi et de reconnaissance est proposée pour la reconnaissance de visages à partir de vidéos.

Lorsque l'on considère la reconnaissance d'objets *génériques* dans les vidéos, les approches rencontrées impliquent comme primitives des *descripteurs locaux* basés sur l'extraction de points d'intérêt, voir par exemple [3, 4, 5, 6]. Ces primitives sont suivies le long de la séquence, de manière à exhiber les plus robustes en sélectionnant celles survivant sur plusieurs trames et en modélisant leur variabilité le long des trajectoires [5]. A noter que ces descripteurs ont été aussi été utilisés pour la reconnaissance d'objets spécifiques, comme les visages, après une étape de détection [7].

En parallèle, avec les récentes propositions de nouveaux descripteurs locaux impliquant différentes natures de support (incluant les patches de texture, les régions homogènes, les formes locales et les points de symétrie), certains travaux [3, 6, 8] ont proposé d'améliorer la description de l'objet pour la reconnaissance dans les images fixes et les vidéos, en exploitant la combinaison de différents descripteurs locaux complémentaires.

Avec le même objectif, d'autres approches ont proposé d'associer un *contexte* plus global aux descripteurs locaux. Dans [9], le vecteur SIFT décrivant chaque point est renforcé par une information de forme dans un voisinage plus large. Dans [10], un contexte (exprimé par des corrélogrammes) est ajouté au descripteur local, intégrant une description des relations spatiales entre le point et ses voisins.

Toutes les approches génériques de description venant d'être décrites sont locales, et tirent ainsi partie des propriétés bien connues de ces descripteurs, comme leur robustesse aux transformations de l'image, aux occultations et aux arrière-plans. Malheureusement, par définition ces primitives ne peuvent pas fournir une description *globale* de l'apparence visuelle de l'objet, pourtant si informative.

D'un autre côté, une description plus globale, comme par exemple la forme ou les couleurs dominantes de l'objet, pourrait être grandement informative pour la reconnaissance. Cela est d'ailleurs démontré dans [11], où la reconnaissance d'organismes marins est améliorée par l'utilisation conjointe de primitives locales et globales. Mais en règle générale, quand les objets sont mêlés à un arrière-

plan, les primitives globales requièrent une étape préliminaire de segmentation de l'image ou de détection de l'objet. Ces traitements sont incompatibles avec la reconnaissance d'objets génériques à partir d'images ordinaires. En effet, si une segmentation peut être facilement réalisée avec des images spécifiques, comme dans [11], elle reste une étape délicate nécessitant des connaissances a priori sur l'image. La détection d'objets requière quant à elle un modèle de l'objet à détecter, par exemple un détecteur de visages [7].

Principe de notre approche. A partir de ces observations, nous proposons une approche de reconnaissance d'objets génériques à partir de primitives visuelles *hétérogènes*. L'objectif est de combiner le potentiel des descripteurs locaux, principalement *leur robustesse* aux occultations et à l'arrière-plan, à celui de descripteurs plus globaux *très informatifs*, sans avoir à considérer un pré-traitement de segmentation de l'image ni de détection de l'objet.
Notre première contribution porte sur la construction *indépendante* de deux espaces de description : l'un associé à la description locale par points d'intérêt, l'autre dédié à la description globale des objets. Dans les expériences menées, nous avons choisi un contour actif comme descripteur global, qui décrit bien la forme de l'objet. Comme énoncé plus haut, les descripteurs locaux permettent la reconnaissance d'objets quel que soit l'arrière-plan, alors que les contours actifs ne sont pas directement utilisables dans un tel contexte. Dans notre approche, les descripteurs locaux sont vus comme la *source primaire* dans une première étape de reconnaissance. Les points appariés obtenus sont vus comme des *ancres* et donnent la possibilité d'aller plus loin dans la reconnaissance. En effet, elles permettent d'*indexer* un ou plusieurs contours actifs, qui viendront alors confirmer ou infirmer la reconnaissance. Pour permettre cela, notre seconde contribution définit des *connexions* entre descripteurs locaux et globaux.
Nous n'avons pas choisi d'enrichir la description locale en ajoutant des composantes aux vecteurs de points, comme dans les approches définissant un contexte de points [9, 10], conduisant à des espaces de grande dimension (188 dans [9] et 960 dimensions dans [10]). Construire séparément les espaces de description et établir des connexions entre eux permet non seulement de garder des dimensions modérées (20 dimensions dans notre cas) restant compatibles avec des applications temps réel, mais également de préserver les avantages de chaque type de descripteur.
L'article est organisé comme suit : dans la section 2, nous présentons les primitives visuelles choisies pour décrire l'apparence des objets. Les descripteurs obtenus sont stockés et structurés dans un catalogue, dont le processus de construction est décrit dans la section 3. Enfin, nous évaluons notre approche de reconnaissance d'objets et démontrons sa pertinence dans la section 4, avant de conclure.

2 Description visuelle des objets

Nous donnons ici les techniques que nous avons employées pour extraire, décrire et suivre les structures locales et globales utilisées comme descripteurs visuels génériques. Ces approches sont classiques, la contribution principale de cet article consiste à structurer efficacement l'espace des descripteurs obtenus et à les utiliser conjointement pour améliorer la reconnaissance.

2.1 Description locale par points intérêt

Les points d'intérêt sont très populaires en vision par ordinateur comme en indexation d'images. Beaucoup d'approches ont été proposées, comme le montre l'étude comparative [12]. Appliquées à la vidéo, il existe aussi des approches temporelles, citons en particulier [13]. Nous avons choisi d'extraire les points d'intérêt trame par trame, puis de les suivre le long de la séquence. L'algorithme d'extraction et de poursuite que nous avons utilisé est similaire à l'algorithme KLT [14].
Pour la reconnaissance, nous caractérisons les points avec les 20 premiers coefficients d'une DCT. Dans le reste de cet article, l'espace des descripteurs locaux sera noté V_n^{point}, ou n est la dimension ($n = 20$). Notons que d'autres descripteurs locaux pourraient être utilisés sans changer le concept de notre approche, en particulier le descripteur SIFT [15], reconnu pour ses performances [12].

Figure 1 – *Exemples de descripteurs locaux et globaux associés à une tête en mouvement.*

2.2 Description globale avec un contour actif

Caractériser l'apparence d'un objet avec des descripteurs de haut niveau peut être effectué avec plusieurs primitives telles que : formes locales, régions, contours actifs, etc. Pour évaluer notre prototype, nous avons choisi de décrire la forme globale de l'objet avec un *contour actif*. Plusieurs raisons nous ont conduit à faire ce choix :
– Un contour actif décrit bien la forme globale de l'objet, et en fournit ainsi une description visuelle très informative (la forme seule de l'objet permet souvent de le reconnaître). Puisque calculés localement, les points d'intérêt ne contiennent pas d'information globale et ne caractérisent pas les mêmes zones de l'objet.
– Un contour actif est assez facile à suivre dans une séquence vidéo ;
– Un contour actif peut aider durant la poursuite des points d'intérêt et vice versa [16].
La théorie des contours actifs fut introduite dans [17], un état de l'art peut être trouvé dans [18]. L'implémentation discrète des contours actifs que nous avons choisie est classique : trois forces sont appliquées à chaque point de contrôle (élongation, courbure et une force externe déduite des contours de l'image), sans connaissance a priori sur

l'objet. L'espace des descripteurs associé est basé sur les descripteurs de Fourier [19]. Dans la suite de cet article, il sera noté V_m^{snake}, où m est sa dimension (ici $m = 20$). A la figure 1, on peut voir un exemple de la caractérisation obtenue pour un objet particulier. Il illustre la complémentarité et la richesse de ces deux descripteurs, puisque la simple vue de ces descripteurs suffit à reconnaître la nature de l'objet sans ambiguïté.

3 Structuration des descripteurs

Nous proposons ici de construire et de structurer un catalogue de caractéristiques visuelles hétérogènes, à partir des catégories de descripteurs venant d'être introduites.

3.1 Construction du catalogue

La figure 2 illustre la structure globale du catalogue. Tous les descripteurs visuels sont collectés lors de la poursuite des objets dans une séquence d'entraînement (voir [A]).

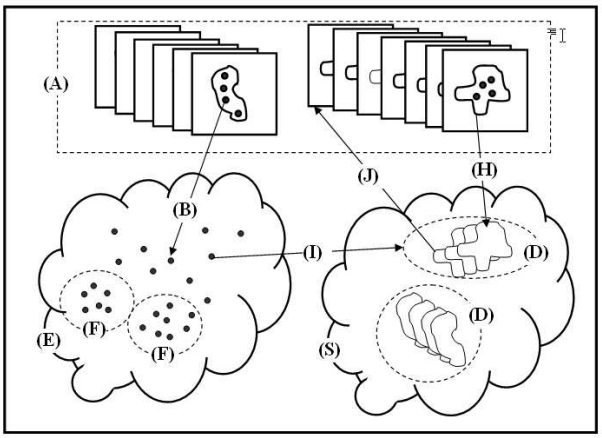

Figure 2 – *Structure globale du catalogue.*

Structuration de l'espace des descripteurs locaux. Pour chaque objet, l'ensemble des descripteurs locaux collectés est inséré [B] dans l'espace V_n^{point}. Chaque point garde le lien avec l'objet associé. Les éléments similaires contenus dans cette espace [E] sont ensuite agglomérés afin de fournir des clusters représentant un vocabulaire visuel (noté " Elementary Local Patterns " ou ELPs) [F]. La construction de vocabulaire visuel à partir de descripteurs locaux a déjà été utilisée dans plusieurs travaux relatifs à la reconnaissance d'objets issus de vidéo, par exemple [5, 6, 7]. Les approches utilisent classiquement un algorithme de regroupement de type k-means, qui fixe le nombre de classes à obtenir et part d'initialisations aléatoires. Dans nos expérimentations, nous avons préféré une approche non supervisée (Competitive Agglomeration - CA) où le nombre de classes est automatiquement déterminé durant le déroulement de l'algorithme, qui est initialisé avec des clusters de points issus de la même trajectoire.

Structuration de l'espace des descripteurs globaux. De façon similaire, tous les descripteurs associés aux contours actifs extraits des objets sont collectés [H] dans l'espace de description V_m^{snake} [S]. Chaque descripteur présent dans V_m^{snake} garde le numéro de trame d'où il est extrait [J]. Tous ces descripteurs de forme sont également soumis à l'algorithme CA [D]. Cette agglomération génère plusieurs clusters que nous appellerons *"Elementary Global Shapes"* ou EGSs.
La structuration des espaces de description V_n^{point} et V_m^{snake} a deux principaux avantages : elle permet une réduction de la redondance spatiale et temporelle des descripteurs, fournissant ainsi des descripteurs plus compacts. Cette compacité a pour conséquence de réduire efficacement le temps de recherche dans ces espaces. Le second avantage est de permettre l'enrichissement dynamique de la description d'objets lors de la reconnaissance d'objets à partir de nouvelles séquences, avec un accroissement minimal du catalogue.

Connexion entre descripteurs locaux et globaux. La dernière étape consiste à *connecter* les espaces de description V_n^{point} et V_m^{snake}. Chaque ELP contient un ensemble de descripteurs locaux similaires et chaque chacun d'entre eux est lié à une trame dans laquelle un contour actif a été extrait et suivi. Ce contour actif est également lié à l'EGS auquel il appartient. Par conséquent, on peut définir une connexion logique entre chaque point de V_n^{point} et un EGS. Les connexions établies entre descripteurs locaux et descripteurs globaux représentent la principale contribution de ce travail : ils sont l'unique façon d'exploiter des descripteurs globaux sans avoir recours à une phase préliminaire de segmentation ou de détection d'objet. Grâce à ces connexions, les descripteurs locaux font fonction d'*index* permettant de déterminer et d'exploiter les descripteurs globaux appropriés.

3.2 Reconnaissance à partir d'une trame

Soit I la trame en cours d'analyse où un objet doit être recherché. La reconnaissance est effectuée en trois étapes :

Étape 1 : Recherche des points candidats. Soit $\{P_1$ ... $P_k\}$ l'ensemble des descripteurs locaux extraits de l'image I. La première étape consiste à rechercher les plus proches voisins de ces points dans l'espace de description V_n^{point}. Pour chaque P_i, les ELPs les plus similaires sont recherchés dans V_n^{point} dans une sphère S_{ϵ,P_i} de rayon ϵ centrée en P_i. Pour tous les P_i considérés, si le nombre de sphères S_{ϵ,P_i} qui intersectent un ou plusieurs ELPs est supérieur à un seuil nommé $T_{anchors}$, alors on suppose que I contient *potentiellement* l'objet (que l'on nomme alors objet candidat). $T_{anchors}$ permet de ne pas détecter un objet systématiquement ; sa valeur est discutée à la section 4.2. Les P_i qui sont appariés dans le catalogue sont nommés M_i ($i \leq k$). D'autres alternatives de classification équivalentes existent pour l'appariement de groupes de points, comme par exemple les classifieurs SVM dédiés aux descripteurs locaux [20, 21]. Nous ne les avons pas utilisées car, à notre connaissance, il n'existe pas de manière efficace de déterminer les M_i à l'issue de la reconnaissance.

Les ELPs du catalogue qui sont appariés avec des M_i sont vues comme des *ancres*. Elle autorisent en effet l'initialisation d'une analyse plus approfondie, car plus globale, des objets candidats, en exploitant les descripteurs globaux EGSs et surtout les connexions établies entre ELPs et EGSs.

Étape 2 : Recherche des formes candidates. Ici, le but est de rechercher dans V_m^{snake} la meilleur forme candidate correspondant à l'objet candidat : l'EGS ayant le plus de connexions avec les ancres est considéré comme le meilleur candidat de forme (on pourrait bien sûr considérer plusieurs candidats de forme). On note SV_{best} le prototype (médoïd) associé au meilleur EGS.

Étape 3 : Validation de la forme candidate. Lors de la dernière étape, il faut vérifier si SV_{best} correspond à l'objet présent dans l'image testée I. Il est nécessaire d'estimer la transformation $\mathcal{T}$ qui existe entre la forme réelle de l'objet dans I et SV_{best}. Puisque SV_{best} est lié à la trame F d'où il vient, nous pouvons apparier l'ensemble des points d'intérêt initialement détectés dans F avec M_i. $\mathcal{T}$ est estimée de ces appariements et permet de placer la forme $SV_I = \mathcal{T}(SV_{best})$ dans I, qui, idéalement, devrait entourer l'objet s'il est présent. Dans le catalogue, chaque point de contrôle pc_i de SV_{best} est décrit par la direction de son gradient $\vec{\nabla} pc_i$. Afin de confirmer si SV_I a une réalité dans I, nous cherchons, dans le voisinage de chacun de ses points de contrôle, si un pixel de l'image possède une direction de gradient proche de $\mathcal{T}(\vec{\nabla} pc_i)$. Si de tels pixels existent pour plus d'un tiers des points de contrôle de SV_I, alors SV_I est déclaré valide, l'objet est déclaré présent dans I et sa localisation est donnée très précisément par SV_I. Le seuil du tiers (noté T_{snake} dans la suite de cet article) a été choisi en fonction de la proportion des occultations autorisées durant la reconnaissance (un exemple de reconnaissance en présence de fortes occultations est montré à la figure 5).
A la reconnaissance de l'objet est associé un taux de confiance $CR(I, O_j)$, où O_j est l'objet reconnu. Ce taux dépend du nombre de points appariés, des distances d'appariement et du nombre de points de contrôle du contour actif qui possèdent un gradient similaire.

4 Evaluation de l'approche

L'approche proposée a été évaluée sur 20 objets aux différentes apparences visuelles en termes de contenu et de forme (des jouets, des visages, des boîtes, etc). Les séquences d'entraînement (au format 352×288 pixels) contiennent chacune 400 trames filmant une rotation complète de l'objet avec un fond uniforme. Nous avons évalué l'approche sur 8000 trames avec les mêmes objets, à une échelle similaire (les descripteurs locaux utilisés n'étant invariants qu'à de faibles changements d'échelle), mais sous des points de vue différents, avec un arrière-plan chargé et une caméra mobile. Dans un premier temps, la reconnaissance est effectuée trame par trame et pour chaque trame, les points intérêt sont extraits de la totalité de l'image. Notre approche est comparée à une approche de référence qui consiste à exploiter seulement les descripteurs locaux pendant la reconnaissance. Les différents résultats de reconnaissance sont présentés et comparés aux sections 4.1, 4.2 et 4.3. Une courbe ROC moyenne y est calculée par objet ; le paramètre de cette courbe est le seuil de détection ϵ de la recherche des plus proches voisins dans V_n^{point} (section 3.2). Puis, dans la section 4.4, nous évaluons notre approche pour un scénario "video-to-video" où la reconnaissance est faite à partir de plusieurs trames. Finalement, la section 4.5 donne une idée des temps de calcul associés.

4.1 Reconnaissance à partir des points seuls

Dans cette évaluation, seuls les descripteurs locaux sont utilisés. Soit $\{P_1 \ ... \ P_k\}$ l'ensemble des descripteurs locaux extraits de la trame I où l'objet est cherché. Les plus proches voisins de chaque P_i sont recherchés dans V_n^{point} à l'intérieur d'une sphère ϵ centrée autour de P_i. Chaque point trouvé vote pour l'objet auquel il est associé. O_j est déclaré présent s'il est associé au vote le plus fort. Remarquons que, comme dans la première phase de la section 3.2 avec $T_{anchors}$, nous imposons un seuil minimal T_j, qui est fonction du nombre moyen N_j de points d'intérêt extraits de toutes les vues O_j de la séquence d'entraînement. Dans cette expérience, $T_j = N_j/3$.
La figure 3 montre plusieurs des 20 courbes ROC obtenues (lignes fines), ainsi que la courbe ROC représentant la moyenne sur les 20 objets (courbe en pointillés épais).

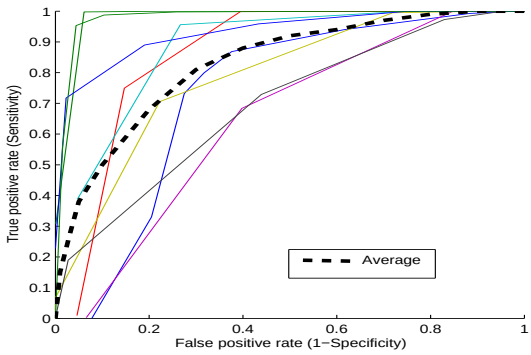

Figure 3 – *Reconnaissance avec les points seuls.*

Le taux d'erreur (*ROC equal error rate*) obtenu est de 74%[1]. Ce résultat a été obtenu en appariant chaque point d'intérêt indépendamment, de sorte qu'il pourrait être amélioré en ajoutant une étape de recalage afin de limiter les mauvais appariements (algorithme Ransac ou Hough par exemple). Ici, il constitue seulement la méthode référence pour évaluer notre approche.

4.2 Apport du descripteur global

La même évaluation a été effectuée en utilisant l'approche complète. Dans la figure 4, la courbe fine en pointillés est celle de la figure 3, présentée ici comme référence et affichée à une échelle adaptée. Les courbes ROC (en fin) illustrent la reconnaissance de plusieurs des objets, alors que

[1] Taux de faux positifs = taux de faux négatifs = 26%.

la courbe plus épaisse représente la courbe ROC moyenne obtenue avec les 20 objets.

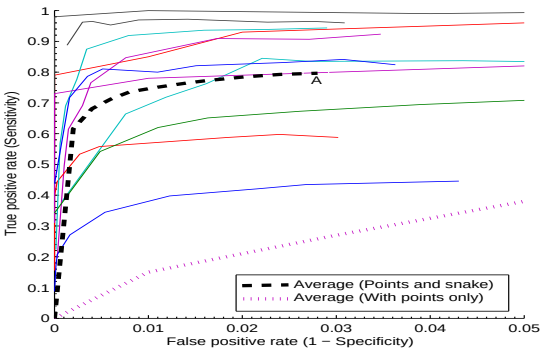

Figure 4 – *Reconnaissance avec les descripteurs globaux.*

Les résultats obtenus sont bien meilleurs lorsque les descripteurs globaux sont utilisés. Par exemple, avec un taux de faux positifs de 2.85 % (point A sur la figure), le taux de faux négatifs est divisé par 3.72 (il est de 74.4% avec l'approche de base et de 20 % avec l'approche complète). Plusieurs raisons expliquent cette amélioration :

Détection automatique et suppression des mauvais appariements. Les points extraits dans l'image et appariés à des vecteurs de V_n^{point} qui ne pointent pas vers l'EGS majoritaire sont déclarés illicites et sont automatiquement supprimés. Ce contrôle efficace et peu coûteux permet de supprimer de nombreux faux appariements.

Détection et suppression des mauvais candidats de forme. Les contours actifs permettent de vérifier la validité de l'objet supposé présent, comme expliqué à la section 3.2. Cette seconde étape de reconnaissance permet de supprimer beaucoup de fausses alarmes, inévitable lorsque les descripteurs locaux sont utilisés seuls.

Ajustement des seuils. Dans les séquences d'entraînement, les diverses vues d'un même objet ne contiennent pas, en moyenne, le même nombre de points. Dans la version basique (sans contour actif), le seuil T_j, représentant le nombre minimum d'ancres requises, est statique pour un objet donné O_j. Dans la version complète, ce seuil (noté $T_{anchors}$ dans la première partie de la section 3.2) peut être ajusté de façon plus fine pour chacune des vues de l'objet. Pour la vue v, $T_{anchors}$ correspond à une fraction α de la moyenne du nombre de points d'intérêt situés dans l'objet 0_j pour l'ensemble des trames qui ont participé à l'EGS associée à v. Cette adaptation contribue à améliorer les résultats de reconnaissance. Dans notre expérimentation, nous avons choisi $\alpha = 1/8$.

La figure 5 illustre un résultat de reconnaissance en présence de fortes occultations. L'importance des occultations tolérées dépend de la valeur des seuils $T_{anchors}$ et T_{snake}. Ici, les résultats ont été obtenus en utilisant les valeurs des seuils données dans l'article.

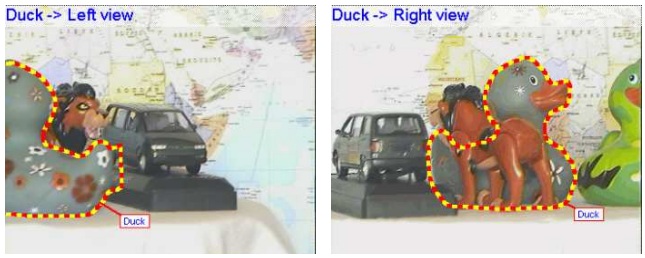

Figure 5 – *Robustesse à de fortes occultations.*

4.3 Choix d'un point de fonctionnement

Pour l'annotation de flux vidéos, il est préférable de choisir un point de fonctionnement (ici ϵ) qui permette de réduire les faux positifs, et donc le nombre d'annotations erronées. Le nombre de vrais positifs s'en retrouvera certainement lui aussi diminué, mais, sous l'hypothèse réaliste qu'un objet est présent sur plusieurs trames d'une séquence, la probabilité de rater cet objet (et donc de ne pas l'annoter) reste faible. En conséquence, le point de fonctionnement que nous avons choisi correspond au cas où le coût des faux positifs est 200 fois plus important que celui des faux négatifs. Les fausses détections correspondantes, déduites des figures 3 et 4, sont résumées dans la table 1.

Approche	Faux Négatifs	Faux Positifs
Points seuls	92%	0.46%
Points + contour actif	38%	0.19%

Table 1 – *Fausses détections pour un point de fonctionnement de rapport 1/200.*

Les résultats obtenus avec l'approche complète démontrent que, lorsque l'objet est présent, il est détecté 15.5 fois par seconde en moyenne (9.5 détections sont manquées chaque seconde, pour un flux vidéo à 25 fps). Cette fréquence de détection (15.5 Hz) permet, de façon quasi certaine, de détecter l'objet s'il passe dans la séquence. Remarquons, qu'avec ce faible taux de faux positifs, nous avons tout de même une fausse détection toutes les 21 secondes en moyenne. Dans la section suivante, nous allons voir qu'en intégrant l'information temporelle, on contribue aussi à réduire le nombre de fausses détections isolées.

4.4 Reconnaissance sur plusieurs trames

Des approches récentes exploitant les flux vidéo proposent d'intégrer les réponses des classifieurs sur plusieurs trames consécutives, comme par exemple [22] qui définit un contexte temporel probabiliste. Pour le moment, nous avons choisi une voie plus simple qui consiste à pondérer les taux de confiance $CR(I_k, O_j)$ (section 3.2) obtenus lorsque l'objet O_j est reconnu dans la trame I_k par les taux de confiance obtenus dans la fenêtre temporelle $[I_{k-w}..I_{k-1}]$ de taille w (dans nos expérimentations, $w = 5$). En intégrant cette information, nous avons amélioré les résultats de la section précédente, voir la table 2.

Scénario	Faux Négatifs	Faux Positifs
mono-trame	38%	0.19%
multi-trames	33.62%	0.069%

Table 2 – *Contribution de l'intégration temporelle avec l'approche complète. Le taux de faux négatifs a été divisé par 1.13 et le taux de faux positifs par 2.74.*

4.5 Temps de calcul

Notre prototype n'est pas entièrement optimisé, mais nous pensons que le temps réel peut être atteint. En particulier, pour le moment aucune structure d'index n'est utilisée pour accélérer la recherche dans V_n^{point}. Les informations suivantes donnent une idée des performances actuelles basées sur un Intel P4 avec une fréquence de 3.2GHz : pour une résolution vidéo de 352×288, avec une moyenne de 100 points d'intérêt extraits dans chaque trame et un catalogue contenant 20 objets (soit environ 240.000 points d'intérêt), le système analyse entre 2 et 3 trames par seconde.

5 Conclusions et perspectives

La principale contribution de ce travail est l'utilisation conjointe de descripteurs locaux et globaux pour la reconnaissance d'objets génériques. Les connexions établies entre eux permettent d'utiliser conjointement des descripteurs locaux *robustes* et un descripteur global *informatif*. Nous avons montré que ces connexions apportent une amélioration significative du processus de reconnaissance dans des vidéos. L'approche ne requiert aucune étape initiale de segmentation d'image ni de détection d'objet afin d'isoler l'objet, la rendant robuste à des arrière-plans quelconques. La mise en place d'espaces de description *indépendants* permet non seulement d'envisager l'annotation temps-réel, de part les dimensions modérées des espaces engendrés, mais aussi de préserver les atouts de chaque technique de description. Finalement, en plus de la richesse de sa description, le choix d'un contour actif comme descripteur global fournit une localisation précise de l'objet retrouvé dans l'image, comme l'illustre la figure 5.

Un pas vers l'annotation fine des objets dans les vidéos. La structure du catalogue autorise *plusieurs niveaux d'annotations* : en plus des labels sémantiques associés aux objets (ici leurs noms), il est en effet possible d'en affecter aux clusters de forme EGSs. Le système a donc la capacité de reconnaître l'objet mais aussi de donner, découlant de la forme reconnue, une idée de sa pose 3D et donc de son comportement dans la séquence. Cette idée est illustrée à la figure 5, avec les annotations "left", "right", "back" et "face" attribuées aux 4 EGSs obtenus pour cet objet.
Nous allons maintenant nous consacrer à l'enrichissement dynamique du catalogue : une fois reconnu, l'objet sera poursuivi dans la séquence, fournissant ainsi de nouveaux descripteurs à ajouter au catalogue. Ces mises à jour permettront de prendre en compte de nouvelles vues de l'objet, rendant la reconnaissance de plus en plus robuste.

Références

[1] F. Rothganger, S. Lazebnik, C. Schmid, et J. Ponce. Segmenting, modeling and matching video clips containing multiple moving objects. Dans *ICCV*, 2004.

[2] S. Zhou, V. Krueger, et R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91 :214–245, 2003.

[3] J. Sivic et A. Zisserman. Video Google : A text retrieval approach to object matching in videos. Dans *ICCV*, 2003.

[4] J. Sivic et A. Zisserman. Video data mining using configurations of viewpoint invariant regions. Dans *IEEE CVPR*, Washington, DC, 2004.

[5] M. Grabner et H. Bischof. Extracting object representation from local feature trajectories. Dans *1st Cognitive Vision Workshop*, 2005.

[6] A. Opelt, J. Sivic, et A. Pinz. Generic object recognition from video data. Dans *1st Cognitive Vision Workshop*, 2005.

[7] J. Sivic, M. Everingham, et A. Zisserman. Person spotting : video shot retrieval for face sets. Dans *CIVR*, 2005.

[8] F. Jurie et C. Schmid. Scale-invariant shape features for recognition of object catgories. Dans *IEEE CVPR*, 2004.

[9] E.N. Mortensen, H. Deng, et L. Shapiro. A SIFT descriptor with global context. Dans *IEEE CVPR*, 2005.

[10] J. Amores, N. Sebe, et P. Radeva. Fast spatial pattern discovery integrating boosting with constellations of contextual descriptors. Dans *IEEE CVPR*, 2005.

[11] D.A. Lisin, M.A. Mattar, et M.B. Blashcko. Combining local and global image features for object class recognition. Dans *IEEE CVPR*, 2005.

[12] K. Mikolajczyk et C. Schmid. A performance evaluation of local descriptors. *IEEE CVPR*, 2003.

[13] I. Laptev et T. Lindeberg. Space-time interest points. Dans *ICCV*, 2003.

[14] C. Tomasi et T. Kanade. Detection and tracking of point features. Technical report CMU-CS-91-132, Avril 1991.

[15] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[16] V. Gouet et B. Lameyre. SAP : a robust approach to track objects in video streams with snakes and points. Dans *BMVC*, Kingston University, London, UK, Septembre 2004.

[17] M. Kass, A. Witkin, et D. Terzopoulos. Snakes : Active contours models. *IJCV*, pages 321–331, 1988.

[18] A. Blake et M. Isard. *Active Contours*. Springer, 1998.

[19] D. S. Zhang et G. Lu.. A comparison of shape retrieval using fourier descriptors and short-time fourier descriptors. Dans *PCM*, pages 855–860, 2001.

[20] C. Wallraven, B. Caputo, et A. Graf. Recognition with local features : the kernel recipe. Dans *ICCV*, 2003.

[21] S. Boughorbel, J.P. Tarel, et N. Boujemaa. The intermediate matching kernel for local image features. Dans *IJCNN*, 2005.

[22] O. Javed, M. Shah, et D. Comaniciu. A probabilistic framework for object recognition in video. Dans *ICIP*, 2004.