Chapitre 4

Classification supervisée des objets vidéo

On va relativiser les "urgences".

Pour créer un film interactif et créer des liens entre des objets "identiques", il faut être capable de repérer les occurrences de ces objets à travers la vidéo. Ce chapitre décrit notre approche de classification semi-automatique des objets suivis, où les classes d'objets d'un film vidéo donné sont connues a priori. Nous appliquons la théorie de mélange gaussien discutée dans le chapitre précédent dans le cas de la recherche automatique des classes d'apparences intra-plan. Cette approche s'applique aussi à la classification discriminante de tous les objets du film dans les classes d'objets pré-définis.

4.1 Introduction

L'étude de la variabilité des objets suivis du chapitre précédent à été menée principalement pour donner des éléments de réponse satisfaisants à la question suivante : comment apparier deux objets suivis ? Les objets suivis étant non-rigides et en mouvements à travers le temps, ils subissent plusieurs changements d'images significatifs, ce qui provoque leur variabilité intra-plan dans l'espace de descripteurs. La généralisation de la méthode des "images-clés" (une image représentative par plan) pour apparier deux objets suivis sur la base de leurs deux occurrences représentatives (occurrence médiane par exemple) parait inacceptable vu la variabilité intra-plan des objets suivis et la difficulté même d'identifier correctement les occurrences d'un même objet suivi (voir l'exemple de la figure 3.2). Capturer la variabilité intra-plan par un mélange gaussien est une meilleure solution pour représenter l'objet suivi; des classes d'apparences intra-plan sont d'abord identifiées et ensuite représentées dans un nouvel espace : l'espace des paramètres gaussiens. En effet, une façon pour apparier deux objets suivis est de prédire les probabilités que n observations provenant de la loi de mélange du deuxième objet suivi soient des réalisations de la loi de mélange du premier objet suivi, ou bien de calculer directement une distance adaptée entre les classes d'apparences intra-plan de ces objets. Il faudra bien sûr accepter que la méthode échoue si les apparences d'un objet dans deux plans ne sont pas similaires. Par exemple, lorsque le même acteur est habillé différemment dans deux endroits du film.

Ainsi, une stratégie de classification des objets suivis se compose d'abord d'une phase de mise en correspondance des apparences de ces objets, et ensuite d'une autre phase de regroupement des objets similaires dans des classes d'objets.

Dans ce chapitre nous proposons une approche de classification supervisée des objets vidéo où on considère les classes d'objets d'un film donné a priori connues. Ici, le constructeur de la vidéo interactive, nommé dans la suite auteur de la vidéo hyperliée, désigne quelques objets différents dans le film vidéo comme "modèles d'objets suivis". Ainsi, chaque modèle d'objet suivi est modélisé d'abord par une loi de mélange gaussien dont le nombre des composantes et le modèle gaussien sont sélectionnés automatiquement en fonction de son degré de variabilité (voir chapitre précédent). Ensuite, la loi globale du mélange de tous les modèles d'objets suivis est construite pour pouvoir classer tout occurrence d'un objet dans le film vidéo, dans la classe d'objets la plus probable a posteriori – règle du maximum a posteriori (classement individuel des apparences). L'étape finale consiste à classer un objet suivi requête dans l'une des classes de modèles d'objets selon le vote majoritaire de toutes ces occurrences classées séparément (classement robuste des apparences).

L'organisation de ce chapitre est le suivant. La section suivante présente un état de l'art des différents types d'applications utilisant une approche de classification supervisée basée sur le mélange gaussien. Les différentes étapes de l'approche proposée sont détaillées dans la section 4.3. Cette approche est validée par des expérimentations réelles sur la séquence vidéo "Avengers-1" qui comporte environ 1391 objets segmentés. Cette séquence est extraite du corpus vidéo de l'INA (voir section 2.5). Une variante de descripteurs globaux de la couleurs est utilisée pour décrire les apparences intra-plan de ces objets. L'évaluation globale des résultats expérimentaux ainsi que l'étude comparative avec l'approche d'indexation classique de "moyenne temporelle de descripteurs" [133] sont détaillées dans la section 4.4. Cette étude montre une augmentation significative de 10% à 35% des pourcentages de bon classement des requêtes par l'approche proposée relativement à l'approche classique. Nous concluons ce chapitre dans la section 4.7.

4.2 État de l'art

Quand il s'agit d'identifier ou de classifier des scènes de visages, plusieurs aspects d'apparences doivent être pris en compte lors de l'étape d'apprentissage du système : visage avec ou sans lunettes, barbe, moustache, face gauche/droite, etc. Sung et Poggio [119] considèrent dans leur système six classes de visages différentes, chacune étant modélisée par une gaussienne. Ils utilisent l'algorithme de centres mobiles pour fabriquer ces classes et la distance de Mahalanobis pour classer les nouveaux visages. Dans un deuxième temps, ils introduisent une classe de rejet (bâtiments, ...) pour raffiner les résultats de détection de visages.

Pour le même objectif, mais cette fois ci sur les séquences d'images, McKenna et al. [82] utilisent le mélange gaussien pour suivre et classifier les visages. Les visages détectés dans les plans sont caractérisés par des distributions de couleurs et parmi ces visages suivis quelques uns sont pris comme modèle et serviront à l'apprentissage du système. Cette approche est la plus proche de celle proposée dans ce chapitre; les visages suivis sont des cas particuliers des objets non rigides suivis dans les plans vidéo; les variations d'un objet quelconque sont beaucoup plus complexes que pour celles associées aux visages. Aussi, plusieurs aspects techniques distinguent notre approche de celle de McKenna et al. et en particulier le choix de la structure de la loi de mélange et les données. Dans [82], les expérimentations ont été menées sur une séquence vidéo fabriquée par les auteurs et sont composées d'une quinzaine de plans. Un taux de bon classement des visages autour de 90% a été obtenu sur cette séquence.

L'utilisation du mélange gaussien commence à s'étendre dans la communauté de vision. Pour le problème de la reconnaissance de gestes, Rosales [104] étudie la performance de plusieurs approches de classification supervisée : K-plus proches voisins, gaussien et mélange gaussien. Chaque action humaine parmi les 7 classes différentes de la base d'apprentissage est caractérisée par des descripteurs simples de mouvements nommés par *Motion History Images* et *Motion Energy Images*, un descripteur étant basé sur la différence des intensités de pixels. Ces expérimentations montrent que le mélange gaussien donne les meilleurs résultats par rapport aux autres méthodes. L'approche de Rosales est en réalité une extension des travaux de Davis [29], où ce dernier utilise les k-plus proches voisins au lieu du mélange gaussien.

4.3 Notre approche

4.3.1 Sélection des modèles d'objets suivis

Les objets localisés dans une séquence vidéo correspondent en réalité à un nombre assez petit de classes d'objets homogènes et distinctes les unes des autres. L'approche que nous proposons dans ce chapitre pour classifier les objets vidéo est semi-automatique. Ainsi, une connaissance a priori d'un échantillon de chaque classe d'objets à fabriquer est fournie. Ici on entend par échantillon un objet suivi (au moins), et donc toutes ses occurrences dans le plan. A cette étape interactive, l'auteur de la vidéo hyperliée intervient pour désigner un modèle d'objet par classe d'objets. Ceci est établi dans le système développé pour ce travail d'une manière interactive et simple [56] : par un simple clic de la souris sur l'image de l'objet d'intérêt, affichée dans la mosaïque de plans de la vidéo (une image représentative par plan), le système enregistre l'objet suivi correspondant comme un modèle d'objet suivi (le système est décrit dans le chapitre 8). La figure 4.1 illustre un exemple de quatre modèles d'objets suivis sélectionnés dans la mosaïque des plans de la séquence vidéo Avengers.

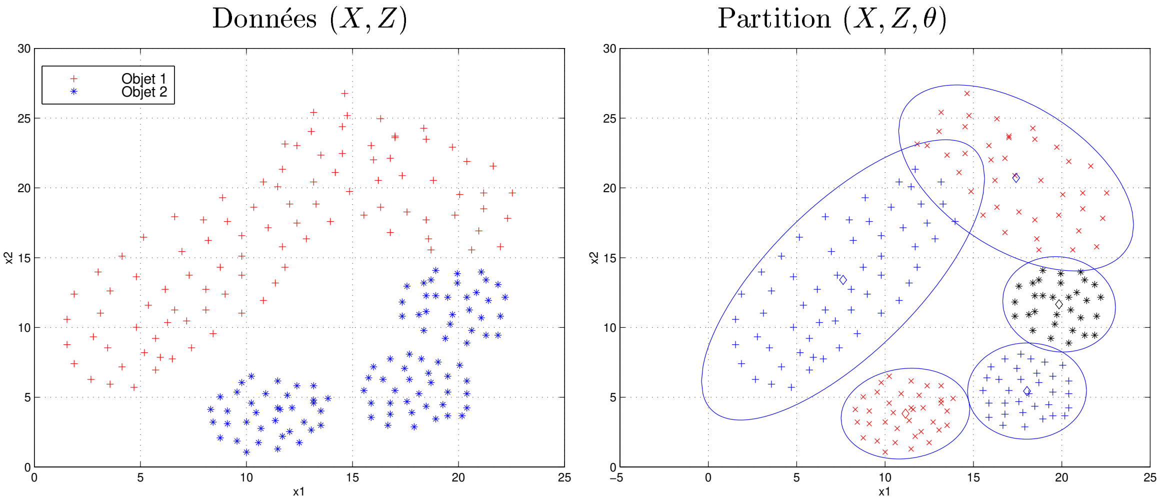
4.3.2 Partition de l'espace de descripteurs

Soit L l'ensemble des modèles d'objets suivis désignés par l'utilisateur expert pour une séquence vidéo donnée. Chaque modèle d'objet suivi représente une classe d'ob-

FIG. 4.1: *Sélection interactive des modèles d'objets suivis dans la mosaïque de plans de la séquence Avengers; l'image médiane de chaque plan est affichée et les contours des modèles d'objets sélectionnés sont aussi dessinés.*

jets; et chaque occurrence est caractérisée par un vecteur descripteur de dimension d, et donc représentée par un point multidimensionnels, $y_i \in \Re^d$. Soit $y = (y_1, ..., y_n)$ les données (observations) collectées de L classes d'objets et $z = (z_1, ..., z_n)$ leurs labels d'appartenance à ces classes: y_i est classé dans la ℓ^{eme} classe d'objets si et seulement si $z_i = \ell$, ($\ell \in \{1, \dots, L\}$). Le nombre d'observations qui appartiennent à la ℓ^{eme} classe (nombre d'occurrences du ℓ^{eme} modèle d'objet suivi) est donné par $n_\ell = \sum_{i=1}^{n}(z_i = \ell)$. La figure 4.2.a illustre les données simulées de deux objets différents dans le plan $\Re^2$. Nous supposons dans la suite que ces observations et leurs labels sont des réalisations indépendamment et identiquement distribuées (i.i.d.) de n couples de variables aléatoires $(\mathbf{Y}, \mathbf{Z}) = \big((Y_1, Z_1), ..., (Y_n, Z_n)\big)$.

La distribution du couple de variables aléatoires $(\mathbf{Y}, \mathbf{Z})$ est un mélange de L distributions (L modèles d'objets suivis). Nous supposons que ces L distributions sont caractérisées



(a) (b)

FIG. 4.2: *Illustration de la distribution simulée de deux objets suivis dans l'espace* $I\!R^2$ *et de leurs classes d'apparences intra-plan déterminées automatiquement. (a) Les données* (X, Z), *avec* $x_i \in \Re^2$, $z_i = \{1, 2\}$, $L = 2$ *et le nombre d'observations de l'objet* ℓ *est* $n_\ell = 100$, $\ell = \{1, 2\}$. *(b) La partition,* θ_ℓ, *qui correspond à chaque objet suivi, déterminée par l'algorithme EM, où, le modèle gaussien général est utilisé, le nombre maximal de composantes gaussiennes (classes d'apparences intra-plan) est fixé à 3, le critère d'information utilisé étant BIC et les nombres de classes retenues étant* $J_1 = 2$ *et* $J_2 = 3$ *pour le premier et le deuxième objet suivi respectivement. Les ellipses de variances sont affichées pour les classes d'apparences intra-plan des deux objets.*

par des fonctions de densité de probabilité $\Phi(y \mid \theta_\ell)$, où $1 \leq \ell \leq L$, et $\theta = (\theta_1, ..., \theta_L)$ représentent les paramètres inconnus de ces distributions qui sont mélangées selon des proportions respectives $p_1, ..., p_L$, avec $p_l = \frac{n_\ell}{n}$ (par exemple proportion libre).

La connaissance du paramètre θ produit une partition de l'espace $I\!R^d$ en L classes, ce qui nous permet alors de **prédire** la classe de toute nouvelle occurrence y d'un objet suivi dans la vidéo (i.e. le vecteur descripteur). Ceci est le principe de la discrimination exclusive où on estime une règle de classement $\hat{u}$ de la forme :

$$\begin{array}{rcl} \hat{u} : I\!R^d & \longrightarrow & \{1, \ldots, L\} \\ y & \rightarrow & \hat{u}(y) \end{array} \tag{4.1}$$

Nous développons dans la suite le principe d'estimation du paramètre θ_ℓ, $1 \leq \ell \leq L$ à partir de données d'apprentissage $(\mathbf{Y}, \mathbf{Z})$ et nous reviendrons après sur la définition de la règle de classement employée par l'approche proposée.

4.3.3 Estimation des paramètres

L'étude détaillée dans le chapitre précédent sur la variabilité intra-plan des descripteurs montre que la distribution d'un objet suivi est souvent multimodale. Une approximation par un mélange gaussien multivarié

$$\Phi(y, \theta_\ell) = \sum_{j=1}^{J_\ell} \rho_j \varphi(y \mid \alpha_j)$$

a été proposée pour capturer cette variabilité. Chaque composante gaussienne, $\varphi(. \mid \alpha_j)$, représente une classe d'apparence intra-plan du l^{eme} modèle d'objet suivi. Elle est caractérisée par un centre μ_j, une dispersion Σ_j autour de ce centre, $\alpha_j = (\mu_j, \Sigma_j)$, et par une proportion ρ_j. L'estimation du paramètre du mélange $\theta_\ell = (\rho_1, \ldots, \rho_j, \alpha_1, \ldots, \alpha_{J_\ell})$ est achevée par la méthode de maximum de vraisemblance dans l'algorithme EM (voir section 3.6.4). Le nombre de classes d'apparences intra-plan n'est pas fixé a priori ni le modèle gaussien qui modélise au mieux la distribution. Ils sont déterminés automatiquement par le moyen des critères d'informations comme BIC et ICL (voir section 3.6.6). Les figures 4.2.a et 4.2.b illustrent respectivement les données simulées de deux objets suivis différents et leurs partitions obtenues dans l'espace $I\!R^2$.

4.3.4 Loi du mélange global

Soient $K = \sum_{\ell=1}^{L} J_\ell$ le nombre total de classes d'apparences intra-plan trouvées pour les L modèles d'objets suivis ($K \geq L$), avec J_ℓ le nombre de composantes gaussiennes retenue automatiquement pour la ℓ^{eme} distribution d'objets suivis, et $\theta = (\theta_1, ...\theta_L)$ leurs paramètres estimés séparément. Seules les proportions sont recalculées en fonction de la nouvelle partition de l'espace $\Re^d$: $p_k = \frac{n_k}{n}$ avec n_k la taille de la k^{eme} classe d'apparence intra-plan et n la taille de données $\mathbf{X}$.

En conséquence le couple de variable aléatoire $(\mathbf{Y}, \mathbf{Z})$ suit une loi du mélange gaussien global de K composantes et de paramètre θ. En un point $y \in I\!R^d$ cette densité du mélange gaussien global est donnée par :

$$\begin{aligned} f(y|z,\theta) &= \sum_{\ell=1}^{L} \frac{n_\ell}{n} \Phi(y, \theta_\ell) \\ &= \sum_{k=1}^{K} p_k \varphi(y|\mu_k, \Sigma_k) \end{aligned} \tag{4.2}$$

Les seuls paramètres à estimer à nouveau sont les proportions de la loi du mélange global.

4.3.5 Classement individuel des apparences d'objets suivis

De là on déduit les probabilités a posteriori que y_i appartienne aux L classes d'objets. Soient Ω_ℓ la ℓ^{eme} classe d'objets (ℓ^{eme} modèle d'objet suivi avec $\ell \in \{1, \ldots, L\}$) et $A_{\ell_1}, \ldots, A_{\ell_{J_\ell}}$, avec $\ell_j \in \{1, \ldots, K\}$, ses classes d'apparences intra-plan qui constituent une partition de Ω_ℓ modélisée par un mélange gaussien de J_ℓ composantes :

Ayant estimé le paramètre θ du mélange global, nous pouvons calculer pour chaque nouvelle occurrence y_i de $I\!R^d$ ($1 \leq i \leq n_r$, avec n_r le nombre total des occurrences de l'objet suivi requête) une probabilité a posteriori qu'elle appartienne à la classe d'apparence intra-plan k ($k \in \{1, \ldots, K\}$) par la formule usuelle de Bayes :

$$t_{ik}(\theta) = Prob(Z = k \mid y_i, \theta) = \frac{p_k \varphi(y_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} p_j \varphi(y_i \mid \mu_j, \Sigma_j)}. \tag{4.3}$$

$$\begin{cases} \forall i \neq j \quad A_{\ell_i} \cap A_{\ell_j} = \emptyset \\ \cup A_{\ell_j} = \Omega_\ell \end{cases} \tag{4.4}$$

les parties A_{ℓ_j} sont disjointes, alors

$$Prob(\Omega_\ell) = \sum_{j=1}^{J_\ell} Prob(A_{\ell_j})$$

et la probabilité a posteriori que y_i appartienne à la classe d'objets ℓ est

$$t_{i\ell}(\theta) = \sum_{j=1}^{J_\ell} Prob(Z = \ell_j \mid y_i, \theta). \tag{4.5}$$

Après le calcul de tous les $t_{i\ell}(\theta)$, $\ell = 1, \ldots, L$, de la même manière que précédemment, on affecte l'occurrence y_i à la classe la plus probable a posteriori : c'est la règle du *maximum a posteriori* connue par MAP (voir figure 4.3). On note dans la suite par *ci*.% le pourcentage de bon classement individuel des apparences d'objets.

Justification du MAP. La théorie Bayésienne de décision justifie la discriminance du MAP en terme de classement. Soit une règle de classement u et soit la fonction $\Im$ de coût 0-1 associant un coût nul aux bons classements et 1 aux mauvais :

$$\begin{cases} \Im(i \mid i)=0 \quad i=1, \ldots, K \\ \Im(i \mid j)=1 \quad i=1, \ldots, K, \ \ i \neq j \end{cases} \tag{4.6}$$

Le risque conditionnel s'écrit

$$\begin{aligned} \mathcal{R}(u \mid y) &= \sum_{k=1}^{K} \Im(u(y) \mid k) t_k(y \mid \theta) \\ &= \sum_{k=1, k \neq u(y)}^{K} t_k(y \mid \theta) \\ &= 1 - t_{u(y)}(y \mid \theta). \end{aligned} \tag{4.7}$$

Moyennant sur tous les x, on obtient le risque moyen qui s'interprète comme une probabilité d'erreur de classement $Prob_e$

$$\begin{aligned} Prob_e(u) &= E_Y[\mathcal{R}(u \mid Y)] \\ &= 1 - E_Y[t_{u(Y)}(Y \mid \theta)]. \end{aligned} \tag{4.8}$$

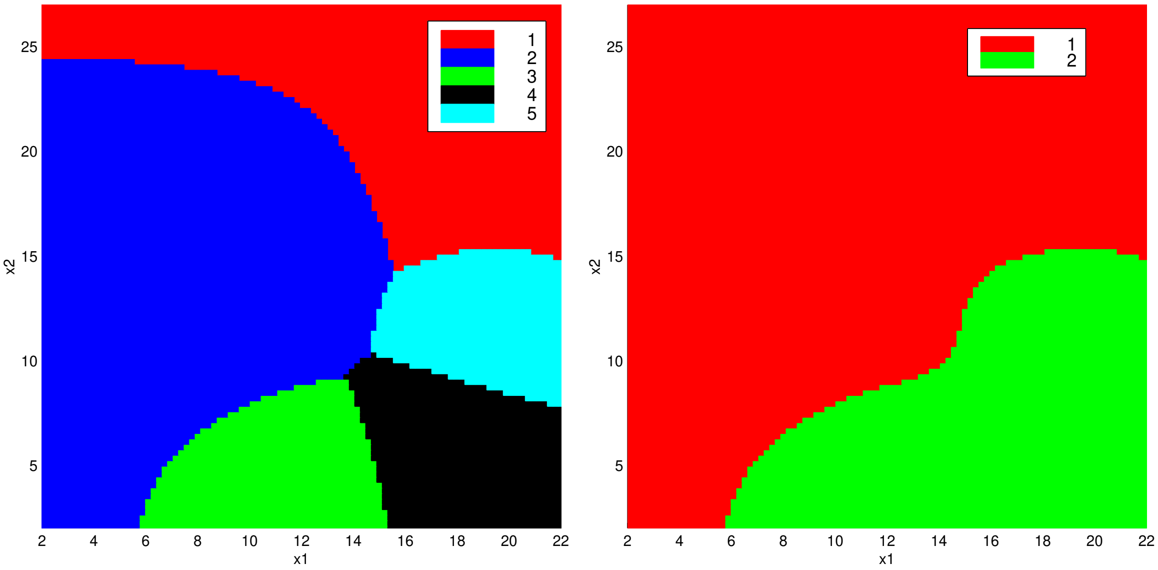


FIG. 4.3: *Illustration de la deuxième étape de l'approche proposée sur les données de la figure 4.2. Classement des points de l'espace* $\mathbb{R}^2$ *par la règle du MAP. (a) Les* 5 *zones d'appartenance aux* 5 *classes d'apparences intra-plan; Pour chaque point de la grille les probabilités a posteriori* $t_{ik}(\theta)$ *d'appartenances aux* $K = 5$ *classes d'apparences intra-plan sont calculées, et en appliquant le MAP, chaque point est classé dans la classe la plus probable; (b) Les zones d'appartenances aux deux classes d'objets d'origine déduites de l'étape précédente (a) par la formule (4.5).*

La règle de classement optimal u_0, ou règle de Bayes, est celle qui minimise l'erreur de classement $Prob_e$. Pratiquement, il suffit de minimiser le risque conditionnel $\mathcal{R}(u \mid y)$ pour chaque y. D'où la règle de MAP

$$\forall y \in \mathbb{R}^d \quad , \quad u_\theta(y) = \arg\min_k t_k(y \mid \theta).$$

Le paramètre $\theta = \hat{\theta}$ étant estimé à partir de données (X, Z), ceci permet d'estimer les probabilités a posteriori par $\hat{t}_{ik}(\theta) = t_{ik}(\hat{\theta})$ et ensuite on en déduit la règle de classement optimale $\hat{u}_\theta(y) = u_{\hat{\theta}}(y)$ (c'est le principe de la méthode du *plug-in*).

4.3.6 Classement robuste des apparences d'objets

Soit Ω_r un objet suivi requête. On cherche à lui trouver la classe la plus proche parmi les L classes d'objets suivis. Le MAP associe chaque occurrence y_i de Ω_r à une classe ℓ. Cependant, tous les y_i ne sont pas probablement associés à la même classe. Ceci est dû principalement, comme nous l'avons déjà mentionné plusieurs fois auparavant, aux variations de l'apparence de l'objet suivi au cours du temps. On se place dans l'hypothèse que seulement quelques occurrences de Ω_r ont des apparences trop éloignées du reste. La figure 3.2 illustre un exemplaire de ce cas : les quelques dernières apparences de l'enfant qui court sont visuellement des apparences aberrantes par rapport à l'ensemble total de ses apparences dans le plan vidéo. Il en résulte alors une mauvaise classification de ces occurrences aberrantes par le MAP.

En appliquant le vote majoritaire, nous rectifions d'une façons robuste les mauvais classements des y_i de Ω_r et ensuite nous identifions la classe de Ω_r. Dans nos expérimentations réelles (section 4.4), cette technique de classement robuste a augmenté le taux de bon classement des occurrences d'objets suivis de 10% ou plus selon les cas. On note dans la suite $cr.$% le pourcentage de bons classements robustes des apparences d'objets.

4.4 Expérimentations

4.4.1 Séquence Avengers-1

Cette séquence de 1016 images comporte 1391 apparences d'objets qui correspondent à 52 différents objets suivis dans 31 plans. Elle est extraite du corpus vidéo fourni par l'INA (section 2.5). Cette séquence servira de base pour les expérimentations de l'approche de classification supervisée proposée dans ce chapitre. Sa particularité est que d'une part plusieurs plans ont des durées très courtes (moins d'une seconde), et d'autre part l'apparence intra-plan des objets est très variable. Toutes sortes de changements de l'image sont présentes dans cette base d'objets : occultation partielle, changement d'éclairage, bruits, changement de point de vue, rotation 3D et changement d'échelle (voir section 3.1.1). Les figures 4.4 et 4.5 illustrent quelques apparences d'objets de cette séquence.

4.4.2 Les modèles d'objets

Un utilisateur expert, peut associer les occurrences de tous les objets suivis de la séquence Avengers-1 à 12 classes d'objets différentes (voiture Ford blanche, voiture Mercedes, acteur J. Steed, actrice Perley, etc). En utilisant le système désigné pour ce travail (voir section 4.3), un modèle d'objet suivi a été sélectionné parmi chaque classe d'objets. De ces 12 modèles d'objets suivis, 448 apparences différentes ont été collectées et serviront pour l'estimation de la loi globale de mélange gaussien. La figure 4.4 illustre 6 différents modèles d'objets suivis; pour chaque modèle d'objet suivi quatres apparences intra-plan sont affichées.

4.4.3 Les données

Le choix d'une caractérisation de bas niveau à stocker sous la forme d'un vecteur de descripteurs, censé représenter une image, est crucial. Un survol des descripteurs des couleurs, des formes et des textures a été présenté dans la section 3.2. Chaque descripteur a ses propres caractéristiques et aucun ne peut être robuste à tous les changements de l'apparence des objets de notre base expérimentale. D'autre part, une restriction de notre approche est l'utilisation des descripteurs globaux uniquement; la modélisation de la variabilité est effectuée dans un espace $\mathbb{R}^d$, où d est la taille du vecteur descripteur caractérisant une apparence d'un objet suivi de la base.

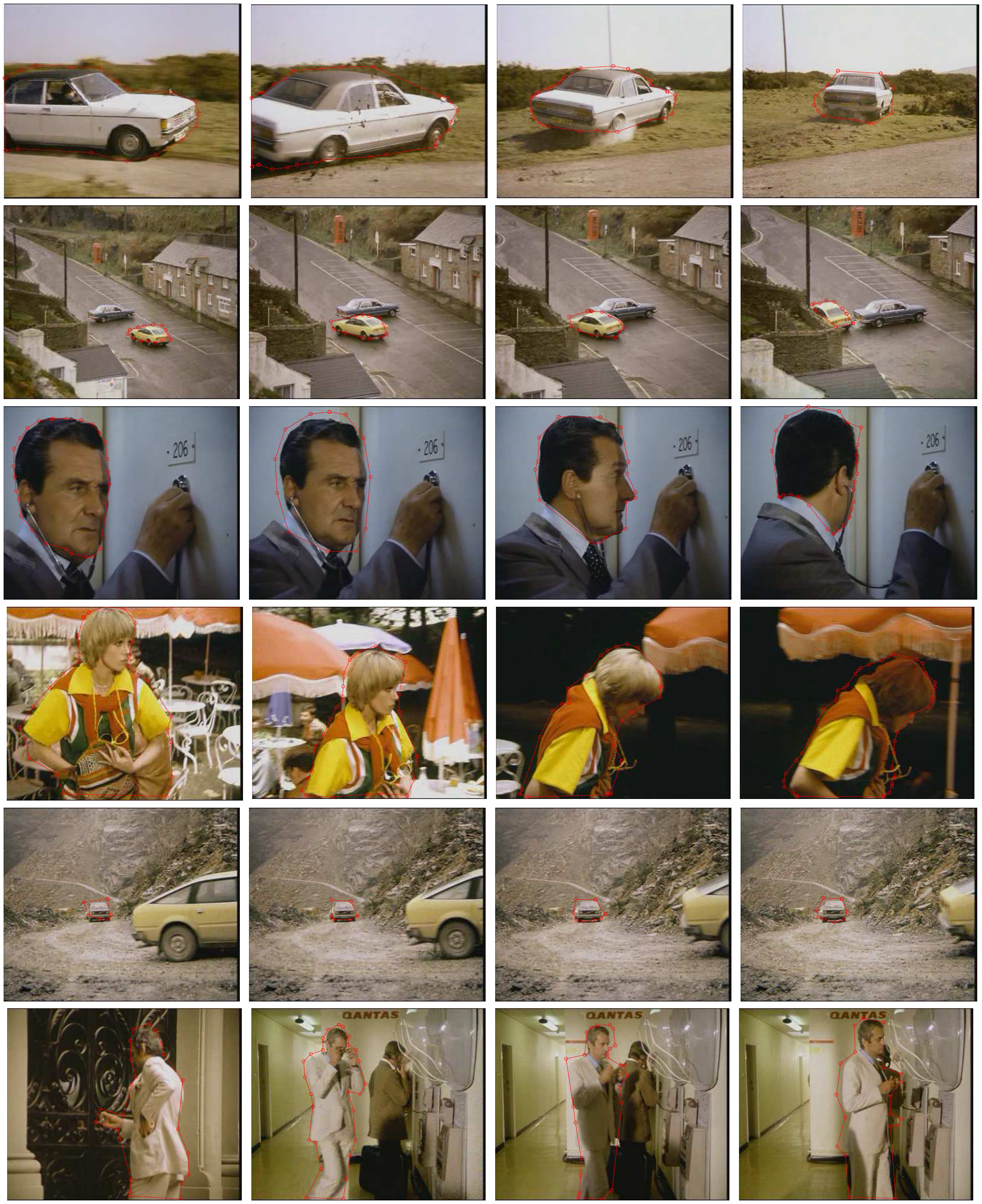


FIG. 4.4: *Quelques modèles d'objets suivis de la séquence vidéo Avengers-1.*

Vu que les objets vidéo sont mobiles et donc leurs apparences à travers le temps sont trop variables, la caractérisation de bas niveaux par des descripteurs de la forme n'est certainement pas le bon choix. Par contre, l'information de la couleur semble être une information visuellement discriminante et aussi très présente dans la base d'objets (voir figures 4.4 et 4.5). Ce choix a été discuté en détail dans la section 3.3.

L'histogramme de couleurs est calculé dans les espaces de couleurs RGB, HSV , HS, H (tinte), S (saturation), I (niveaux de gris), rgb (chromaticité) et $l_1l_2l_3$. Afin de prendre en compte le nombre réduit des apparences des modèles d'objets suivis, les espaces de couleurs sont quantifiés en un nombre de couleurs raisonnable mais suffisament discriminant : les espaces RGB, HSV, rgb, $l_1l_2l_3$, HS, H, S, et I sont quantifiés en 64, 64, 64, 49, 32, 32, 32 et 32 couleurs respectivement. En plus, on considère que les axes d'un histogramme disposent d'un nombre de cellules équivalent (par exemple 4 cellules sur les 3 axes R, G et B de l'espace RGB). Notons que la quantification de l'espace de couleurs est faite d'une façon empirique et que d'autres dimensions peuvent être considérées [127]. L'étude de l'effet de la variation de la dimension de l'histogramme sur l'appariement d'images sort du cadre de ce travail.

Pour chaque apparence de la base des objets suivis (de toute la séquence vidéo) on extrait les histogrammes cités ci-dessus. Ensuite, sur chaque tableau de données $n \times d$ ainsi obtenu pour les n apparences d'objets de la séquence vidéo traitée, où chaque ligne de ce tableau est un histogramme de d dimensions, l'analyse en composante principale est appliquée (voir section 3.4).

Ceci permet de projeter les histogrammes RGB, HSV, rgb, $l_1l_2l_3$, HS, H, S et I de la séquence Avengers-1 dans les espaces à $10, 10, 3, 10, 8, 5, 8$ et 8 dimensions respectivement. Une qualité de représentation Q_E (formule 3.5) des données dans ces espaces réduits de 95% est fixée a priori. Dans la suite, on note par d et d_E les dimensions de l'espace de descripteurs initial et réduit respectivement.

4.4.4 Paramètrage de l'approche

Dans un premier temps, les classes d'apparences intra-plan de chaque modèle d'objet suivi sont identifiées. Ces classes sont obtenues par une modélisation de la variabilité de l'objet suivi dans l'espace de descripteurs $\mathbb{R}^{d_E}$, par un mélange gaussien. Le nombre de composantes gaussiennes maximal, noté $MaxNbC$, peut varier de 1 à 4. Trois critères, BIC, ICL et NEC sont utilisés dans les expérimentations pour sélectionner automatiquement la meilleure structure de la loi de mélange (voir section 3.6.6).

Les 12 lois de mélanges des modèles d'objets suivis sont ensuite réunies sous la forme d'une seule loi globale de mélange gaussien.

4.4.5 Les requêtes

La loi globale de mélange gaussien ainsi obtenue pour les 12 modèles d'objets suivis permet au système de classer toute nouvelle apparence d'un objet suivi (requête individuelle) dans la vidéo Avengers-1 dans une classe unique (voir section 4.3.5). Dans un premier test, 943 nouvelles apparences différentes qui ne sont pas auparavant utilisées

Fig. 4.5: *Quelques apparences requêtes de la séquence Avengers-1.*

dans l'apprentissage de la loi de mélange global sont classées par le système. La figure 4.5 illustre quelques apparences d'objets requêtes. Aussi, les 448 apparences utilisées dans l'apprentissage sont aussi classées, afin de tester l'effet de la fusion des lois de mélanges des modèles d'objets suivis dans une même loi globale. Rappelons que dans ce cas chaque apparence d'un objet suivi est classée indépendamment des autres apparences du même objet suivi. Dans un deuxième test, les 52 objets suivis sont considérés eux même comme des requêtes où toutes les apparences d'un même objet suivi sont rangées dans la même classe : approche de classement robuste (voir section 4.3.6).

4.4.6 Les résultats

Une comparaison avec les classements manuels basés sur la vision humaine semble être le seul moyen pour évaluer les résultats de l'approche proposée. En adoptant cette démarche, les pourcentages $ci.\%$ et $cr.\%$ de bon classement individuel et robuste sont comptés pour la totalité des requêtes (voir section 4.3.5 et 4.3.6).

Les résultats de classement sont rangés dans les tableaux 4.1 et 4.2. Le premier rang indique l'espace de descripteurs sur lequel la loi globale de mélange gaussien est construite. Les dimensions initiales et réduites de cet espace sont indiquées dans le deuxième et troisième rang respectivement. Les pourcentages $ci.\%$ et $cr.\%$ sont montrés dans le cas où les critères BIC, ICL et NEC sont utilisés. Une illustration graphique de ces tableaux est donnée dans les figures 4.8 et 4.9.

L'analyse de ces tableaux permet de tirer plusieurs conclusions très intéressantes des différents aspects de l'approche :

- Les **meilleurs pourcentages** obtenus avec les approches de classement individuels et robustes des apparences d'objets sont de 81.5% et 90.2% respectivement. La modélisation de la variabilité intra-plan des objets suivis est effectuée dans ce cas dans l'espace de descripteurs RGB à 10 dimensions, où le nombre maximal de classes d'apparences intra-plan ($MaxNbC$) est égal à 2 et où les critères BIC et ICL sont employés. Ces scores sont très satisfaisants vu la particularité de la base d'objets de la séquence Avengers-1 (voir section 4.4.1). Les figures 4.6 et 4.7 montrent quelques résultats de classement des requêtes issues de cette expérience.

- "**Mieux** la variabilité est modélisée – en terme du nombre de composantes gaussiennes et de la complexité du modèle gaussien sélectionné – **mieux** les frontières des classes d'objets sont estimées et donc **plus** les classements sont corrects". Cette règle est validée expérimentalement, dans presque tous les espaces de descripteurs testés, lorsque le nombre maximum de composantes gaussiennes ($MaxNbC$) progresse de 1 à 2; ainsi on observe une amélioration des pourcentages $ci.\%$ et $cr.\%$ entre 20% (exemple de $cr.\%$ dans l'espace $l_1l_2l_3$ avec BIC) et 4%. Par contre, ces résultats commencent à se dégrader quand $MaxNbC$ dépasse 2, dans les espaces de descripteurs RGB, HSV et $l_1l_2l_3$, ce qui n'est pas étonnant car le nombre des individus qui ont participé à l'apprentissage n'est pas assez suffisant pour estimer des frontières optimales entres les classes d'objets : la distribution d'un objet suivi parait multimodale, ainsi avec un nombre des apparences assez petit par rapport à la taille

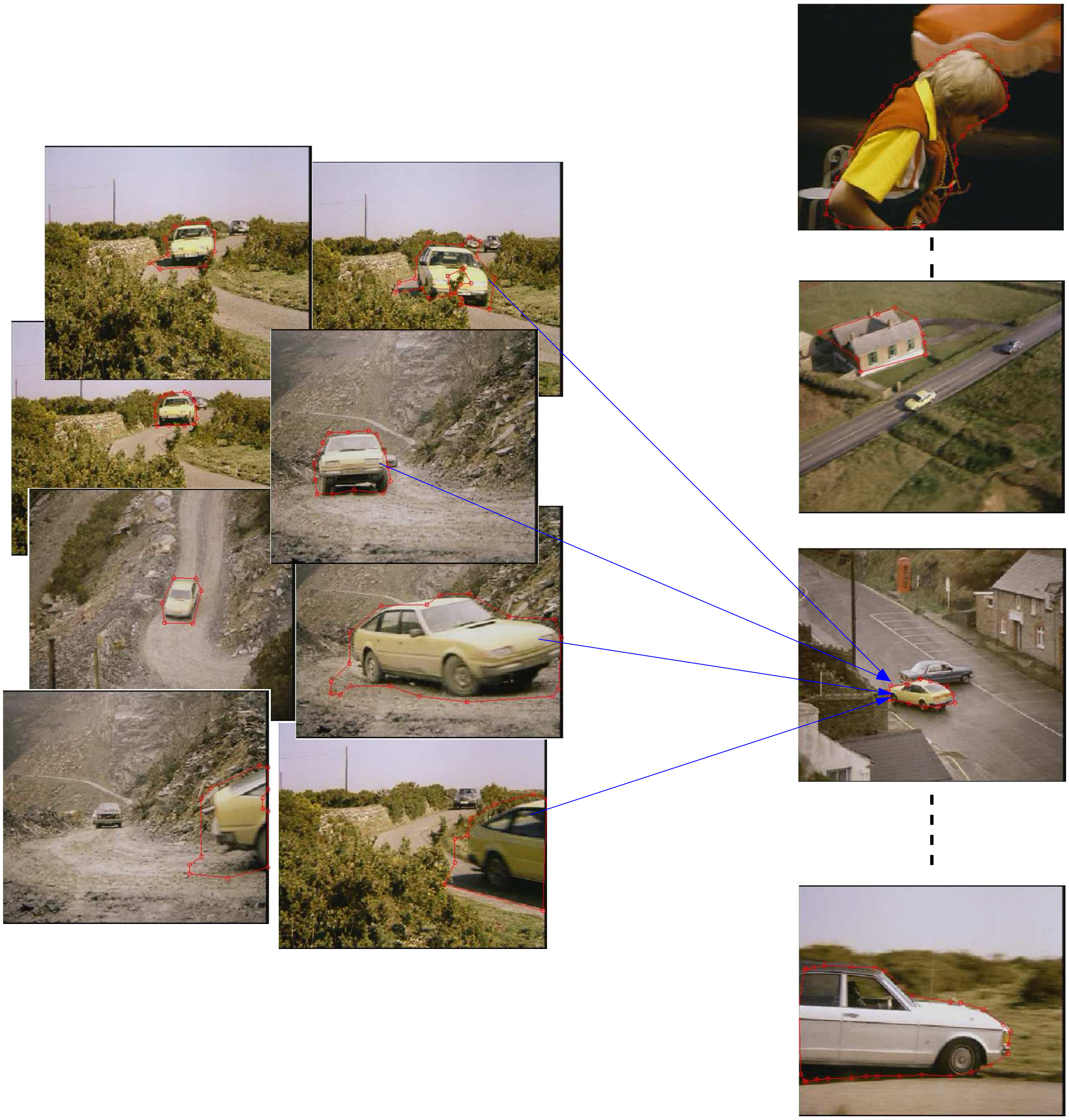


FIG. 4.6: *Résultats de classement de quelques apparences requêtes dans les modèles d'objets suivis de la séquence vidéo Avengers-1. Toutes ces requêtes sont correctement classées.*

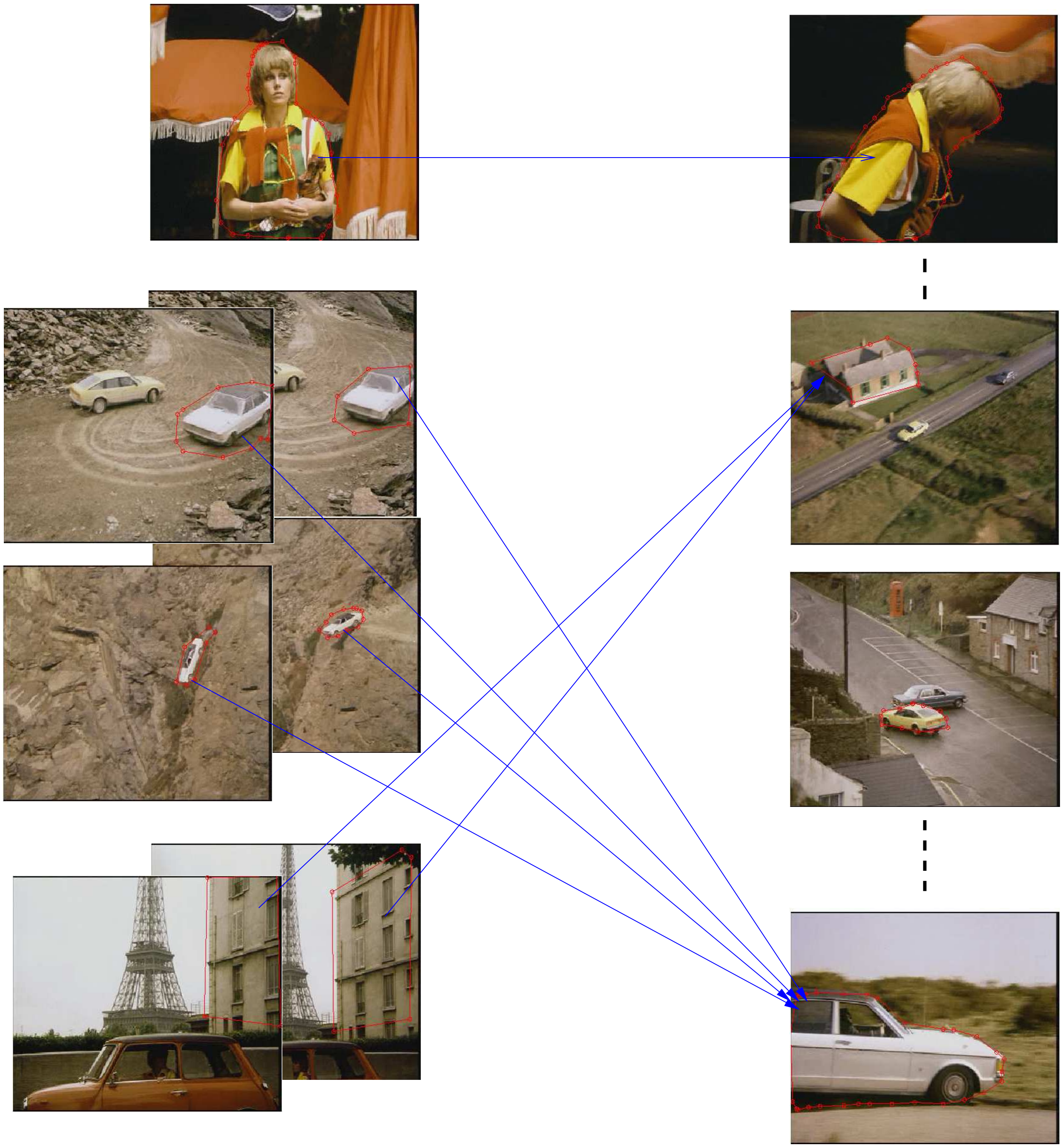


FIG. 4.7: *Résultats de classement de quelques apparences requêtes dans les modèles d'objets suivis de la séquence vidéo Avengers-1. Toutes ces requêtes sont correctement classées.*

de ces espaces les critères utilisés ont souvent tendance à choisir un nombre de classes assez grand, mais avec des modèles simples. Dans des espaces de dimensions plus petites la validité de la règle ci-dessus est manifestement plus nette. Par exemple, dans les espace HS, H, S et I à 8 dimensions, on trouve que les pourcentages de bons classements sont les plus élevés quand $MaxNbC$ est égal à 4. On a *cr.%* est égal à 71.9%, 63.3% et 61.8% dans les espaces HS, I et H respectivement. Le critère ICL a été utilisé dans ce cas.

Espace de descripteurs	d	d_E	$MaxNbC$	*ci.%*			*cr.%*		
				BIC	ICL	NEC	BIC	ICL	NEC
RGB	64	10	1	77.9	77.9	43.7	85.3	85.3	46.8
RGB	-	-	2	81.0	81.5	57.9	90.2	89.1	58.9
RGB	-	-	3	81.0	80.3	65.7	86.5	79.9	69.3
RGB	-	-	4	77.2	79.9	57.8	82.5	86.8	62.3
HSV	64	10	1	67.6	67.6	44.4	68.3	68.3	38.9
HSV	-	-	2	70.2	67.4	41.6	77.8	72.3	39.1
HSV	-	-	3	66.5	69.1	43.5	71.7	75.8	43.7
HSV	-	-	4	66.4	66.7	59.2	68.6	71.7	56.7
HS	49	8	1	60.2	60.2	44.2	54.0	54.4	44.6
HS	-	-	2	55.7	55.6	50.4	64.0	59.2	45.1
HS	-	-	3	53.6	52.0	68.6	55.0	52.9	67.6
HS	-	-	4	52.6	67.0	63.5	50.3	71.9	62.4
I	32	8	1	42.8	42.8	42.8	45.4	45.8	45.4
I	-	-	2	49.4	51.0	48.7	52.0	51.0	48.6
I	-	-	3	52.0	51.2	51.7	52.4	51.0	55.4
I	-	-	4	52.0	52.0	48.6	54.1	63.3	51.0

TAB. 4.1: *Résultats des approches de classement individuels et robustes des apparences d'objets de la séquence vidéo Avengers-1, dans les espaces de descripteurs RGB, HSV, HS et I.*

4.5 Approches classiques de reconnaissance des objets vidéo

Une approche classique d'indexation des objets d'une séquence vidéo consiste à indexer seulement les objets qui se trouvent dans les images clés [92]. Cette approche est considérée comme une solution simplificatrice du problème de gestion de la taille gigantesque de la vidéo (voir l'introduction du chapitre précédent), en ignorant le problème de la variation intra-plan du contenu. Notons que cette approche a été implémentée dans la première version de notre système de construction de la vidéo hyperliée (voir chapitre 8). L'image médiane est souvent prise comme image représentative du plan. Plus tard, Zhang [133] propose la méthode de "moyenne temporelle de descripteurs" (appelé *temporal mean*

Espace de descripteurs	d	d_E	*MaxNbC*	*ci.%*			*cr.%*		
				BIC	*ICL*	*NEC*	*BIC*	*ICL*	*NEC*
H	32	5	1	54.5	54.5	38.0	54.7	54.7	43.1
H	-	-	2	60.5	54.7	59.2	61.8	52.1	65.8
H	-	-	3	57.1	50.5	45.1	59.2	50.7	45.9
H	-	-	4	58.2	60.5	44.8	62.9	61.8	47.4
S	32	8	1	48.5	48.5	39.5	47.1	47.1	39.0
S	-	-	2	59.5	50.8	50.8	63.0	56.6	50.2
S	-	-	3	45.1	51.3	48.8	45.6	55.9	51.3
S	-	-	4	53.1	49.5	46.2	52.6	55.2	45.4
rgb	64	3	1	36.2	36.2	33.8	40.3	40.3	36.7
rgb	-	-	2	38.5	37.2	34.8	38.5	40.0	39.3
rgb	-	-	3	40.5	37.4	35.1	41.4	41.3	36.7
rgb	-	-	4	38.1	38.6	36.0	40.8	41.7	38.9
$l_1l_2l_3$	64	10	1	43.0	43.0	39.8	41.8	41.8	42.3
$l_1l_2l_3$	-	-	2	56.6	54.6	38.2	61.3	61.5	38.0
$l_1l_2l_3$	-	-	3	55.0	55.8	43.3	58.7	60.5	46.2
$l_1l_2l_3$	-	-	4	51.3	52.4	41.0	55.6	53.7	41.3

TAB. 4.2: *Résultats des approches de classement individuels et robustes des apparences d'objets de la séquence vidéo Avengers-1, dans les espaces de descripteurs H, S, rgb et* $l_1l_2l_3$.

feature en anglais) qui intègre l'aspect temporel dans le calcul de la similarité entre les plans vidéo; il calcule la moyenne de la totalité des descripteurs extraits des images du plan, en particulier il extrait la luminosité et quelques couleurs dominantes dans l'image.

Avec comme objectif la comparaison de performance de l'approche proposée et la méthode de Zhang, nous avons testé cette dernière sur la base d'objets de la séquence Avengers-1. Effectivement, la méthode de Zhang est un cas particulier de notre approche, où chaque modèle d'objet suivi est représenté par le centroïde μ de la distribution avec une variance nulle, commune pour tous les modèles d'objets suivis. Dans la base d'indexes, il y aura donc seulement 12 descripteurs. Le jeu de requête utilisé auparavant est considéré dans l'évaluation de cette approche classique. D'abord, une requête est appariée avec la base par le calcul d'une distance euclidienne. Ensuite, elle est classée par la méthode du premier proche voisin. Et enfin, un classement robuste est appliqué. Le tableau 4.3 décrit les résultats de test de cette approche, dans le cas de classement individuel et robuste, et ceci pour tous les espaces de descripteurs à dimensions réduites.

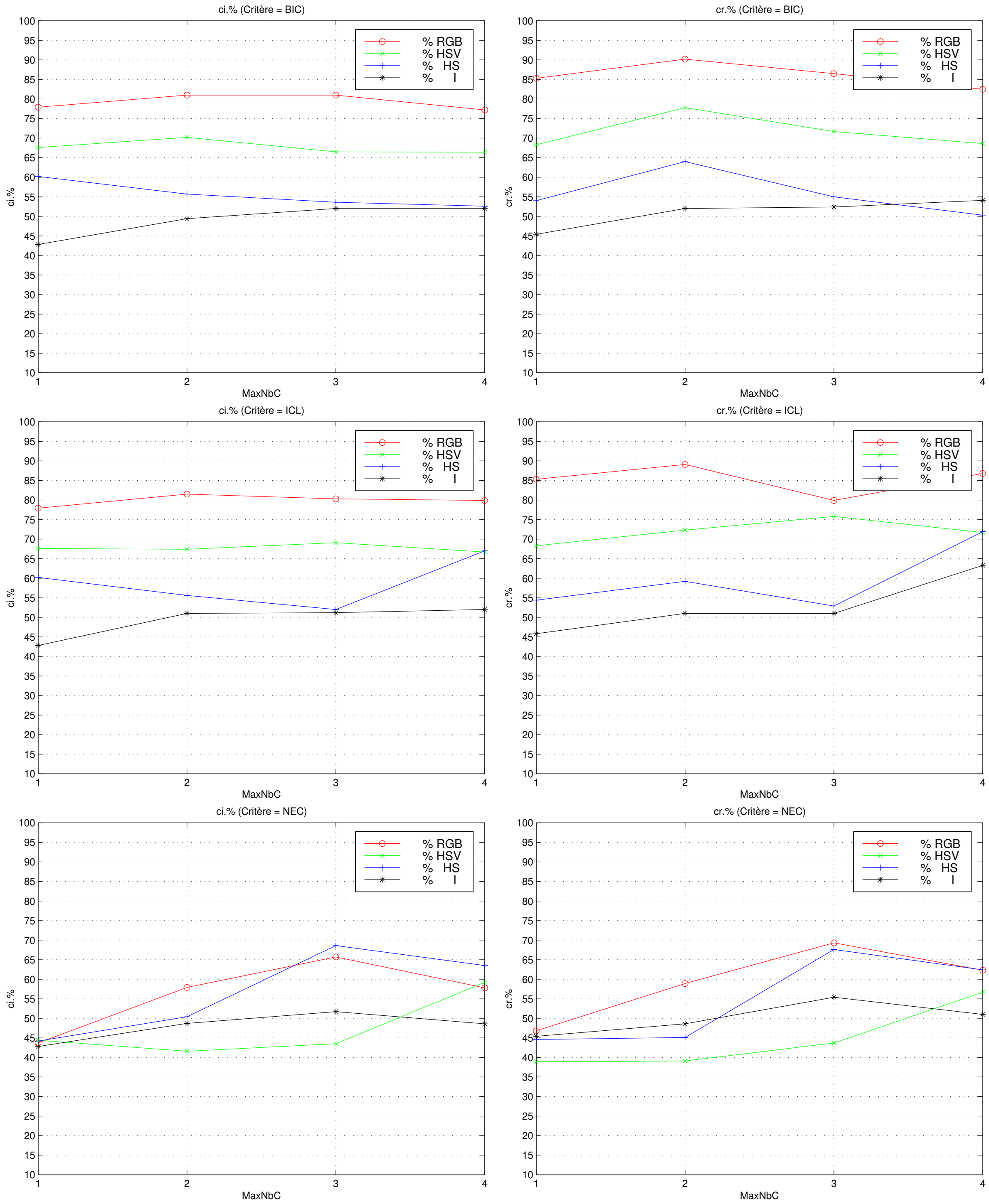


FIG. 4.8: *Illustration graphique des résultats des approches de classement individuels (colonne gauche) et robustes (colonne droite) des apparences d'objets de la séquence vidéo Avengers-1; Les résultats avec les critères BIC, ICL et NEC sont affichés dans le premier, deuxième et dernier rang respectivement.*

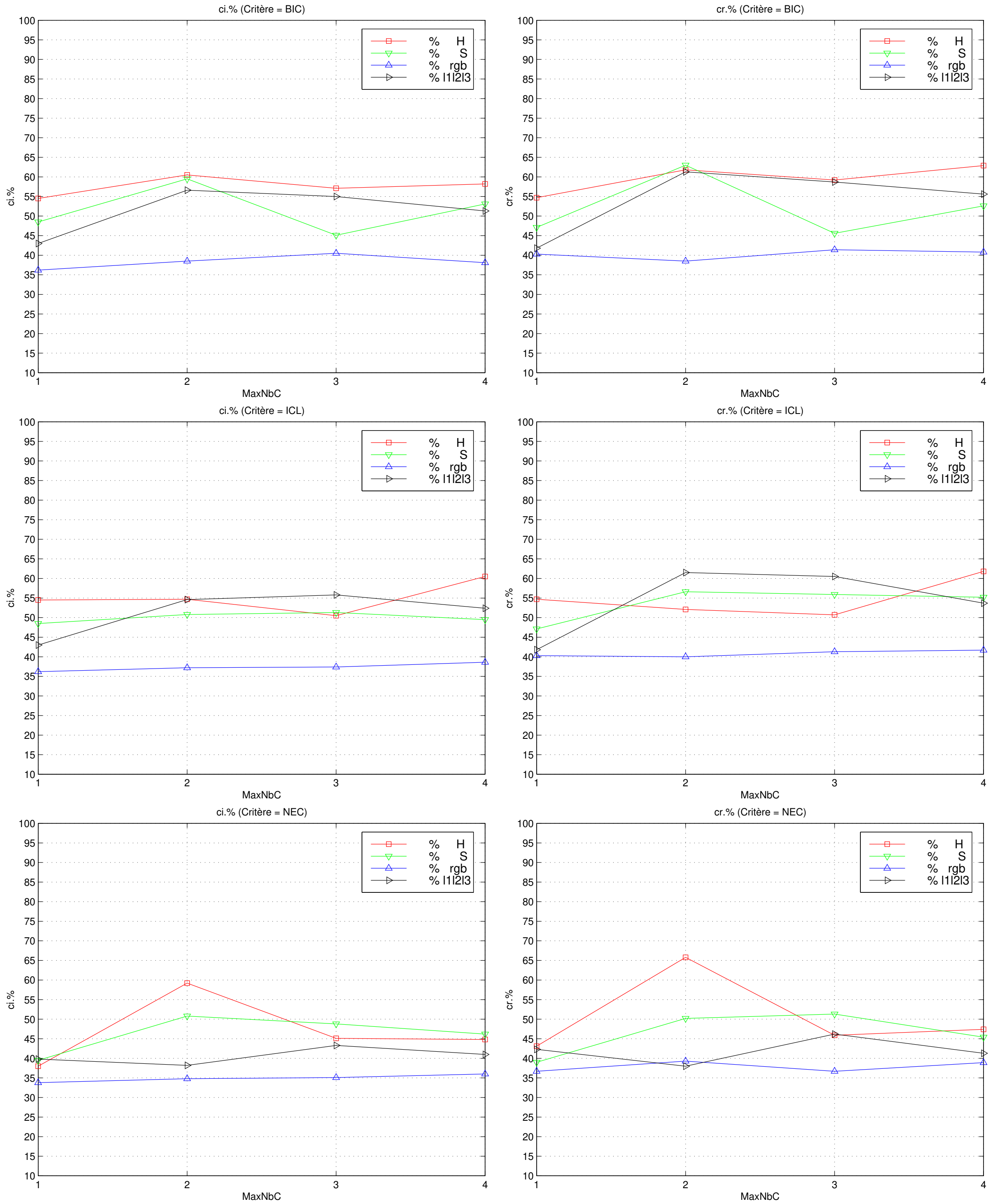


FIG. 4.9: *Suite de la figure 4.8*

Descripteur	d_E	*ci.%*	*cr.%*
RGB	10	56.6	55.3
HSV	10	53.7	60.3
HS	8	48.2	48.4
I	8	36.6	41.1
H	5	42.4	48.7
S	8	39.5	39.5
rgb	3	35.4	41.3
$l_1 l_2 l_3$	10	42.4	42.7

TAB. 4.3: *Résultats de scores de bon classement par l'approche classique de la "moyenne temporelle de descripteurs".*

4.6 Analyse comparative et discussion

Performance de l'approche proposée. L'approche de classification supervisée proposée dans ce chapitre pour classifier les objets suivis en groupes homogènes est basée sur une idée fondamentale : prendre en compte la variation de l'apparence intra-plan des modèles d'objets suivis. L'intégration de cet aspect temporel consiste à identifier les classes d'apparences intra-plan des objets suivis indépendamment par des lois de mélanges gaussiens. Le nombre de classes ainsi que le modèle gaussien adaptés à une distribution d'un objet suivi dans l'espace de descripteurs sont déterminés à l'aide des critères probabilistes. Même si l'objet suivi n'est pas mobile dans le plan vidéo, souvent sa distribution dans l'espace de descripteurs n'est pas compacte autour d'un centre et avec une variance (presque) nulle pour les raisons citées dans l'introduction de ce chapitre. Il en résulte que la méthode de Zhang (*moyenne temporelle de descripteurs*) semble être inefficace pour indexer les objets vidéo. Par contre, notre approche est en quelque sorte une extension de cette approche classique où une distribution est représentée par un ou plusieurs centres avec des variances autour de ces centres. On s'attend donc à ce qu'elle soit beaucoup plus performante; les expérimentations ci-dessus montrent une amélioration significative des pourcentages de classements robustes de 10% à 35% par notre approche. La figure 4.10 illustre les performances de deux approches.

Classement robuste. L'approche de classement robuste améliore les résultats de classement individuel des apparences d'objets jusqu'à 10%. Ceci est un peu faible. Lorsqu'un peu d'apparences d'un objet suivi sont aberrantes et que la majorité est bien classée, alors le classement robuste rectifie les mauvais classements. Dans ce cas, les pourcentages de bon classements vont progresser et, dans le cas inverse, ils vont décroître. Sur la base d'objets ainsi expérimentée, ces deux situations sont malheureusement présentes, l'une pénalisant l'autre. Par contre, le premier cas est plus présent car les pourcentages ont augmenté.

Le deuxième cas, où la majorité des apparences d'un objet suivi est mal classée, peut

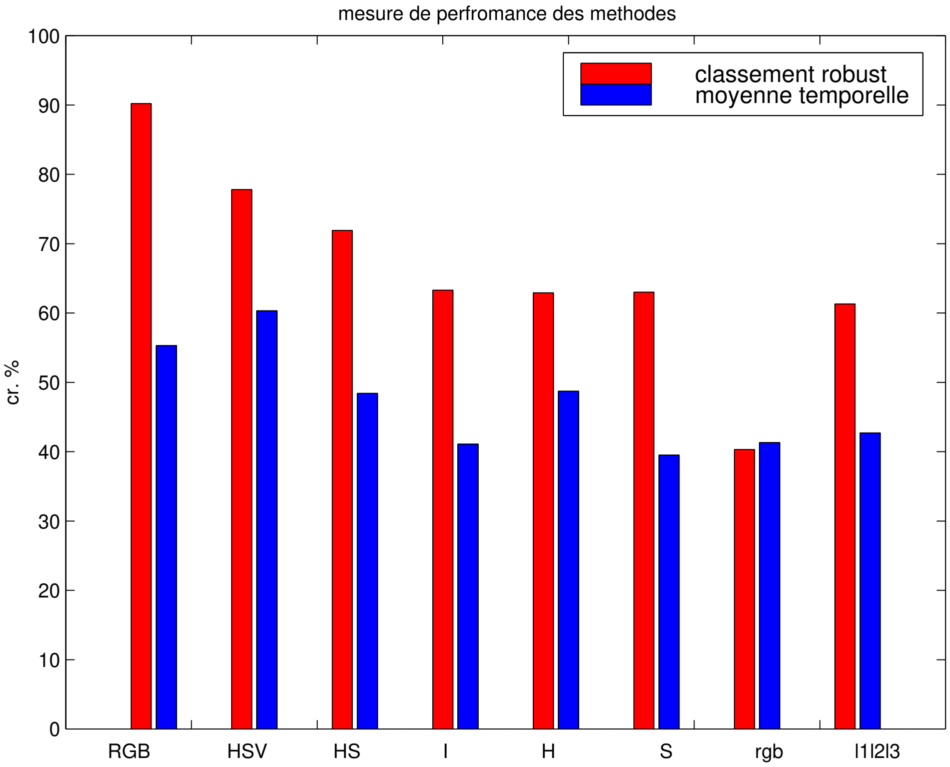


FIG. 4.10: *Comparaison des résultats de l'approche de classement robuste et celui de la "moyenne temporelle de descripteurs".*

être expliqué par le fait que ces apparences sont trop éloignées (visuellement et dans l'espace de descripteurs) du modèle d'objet suivi. Notons que, dans ce travail, on suppose que pour chaque requête existe un modèle d'objet. Par contre, un modèle d'objet suivi est sélectionné dans le plan vidéo et parfois ses apparences ne recouvrent entièrement pas toutes les apparences dans les autres plans. Il serait en de posséder d'un modèle 3D du modèle d'objet d'où l'on pourrait produire des apparences virtuelles multiples à savoir le modèle convenable de transformation du modèle d'objet. Cela permettrait de résoudre le problème des données limitées pour l'estimation.

D'autre part, si les modèles d'objets suivis ne sont pas représentatifs (aux moins dans le sens sémantique) de toutes les requêtes, un processus de rejet est indispensable pour le système. On peut utiliser la loi de χ^2 pour rejeter toute apparence déjà classée dans k par la règle de MAP, qui a une distance de Mahalanobis à cette classe k qui n'est pas dans un intervalle de confiance de 95% par exemple. Du point de vu technique, le rejet n'a aucun sens; des objets dans la vidéo ne font pas partie des classes d'objets ainsi fabriquées par le système dont l'objectif est la création de la vidéo hyperliée : un objet point sur toutes ces occurrences dans la vidéo.

Notons que le changement de la base des modèles d'objets suivis peut influencer relativement les résultats. Dans [57], nous avons changé légèrement la base d'objets en utilisant

15 modèles d'objets suivis et des changements mineurs des résultats ont été remarqués.

Structure de la loi de mélange. Le choix d'un critère pour déterminer la structure d'un mélange gaussien a un effet significatif sur les résultats finaux de classement des requêtes. Lorsque la distribution de chaque objet suivi est modélisée par une seule composante gaussienne ($MaxNbC = 1$) le critère NEC semble être mal adapté pour le choix du modèle gaussien le plus discriminant (frontière optimale) parmi les 7 modèles gaussiens mis en compétition. C'est la seule explication des mauvais résultats obtenus lorsque ce critère est employé. Par contre BIC et ICL donnent des meilleurs résultats par la sélection du modèle gaussien le plus complexe.

Les tests prouvent que dans les espaces de descripteurs RGB, HSV, $l_1l_2l_3$ et rgb la modélisation des distributions des modèles d'objets suivis par des lois de mélanges de deux composantes gaussiennes au maximum donne les meilleurs résultats. Ceci dit que avec un nombre de composantes plus élevé le risque de mal adaptation (*over-fitting* en anglais) des données était aussi très grand; l'estimation des paramètres gaussiens est devenue instable car le nombre des individus par classe gaussienne est trop petit par rapport à la dimension de l'espace d'apprentissage. Par contre, dans les espaces de dimension plus petite I, H, S, et HS, la modélisation avec 4 gaussiennes au maximum donne les meilleurs résultats de classement. Le critère ICL semble être le mieux adapté dans ce cas.

Descripteurs invariants aux changements de luminosité. La méthode de classement fournit des très bons résultats dans l'espace de descripteurs RGB et de très mauvais résultats dans l'espace rgb normalisé. Idem, avec la méthode classique de Zhang. C'est surprenant vis à vis des résultats obtenus par Finlayson [36] et Gevers [43]. Mais, deux choses expliquent ce phénomène : (1) un histogramme rgb quantifié en 64 couleurs contient plus de deux tiers de cases vides (fréquences nulles), car les couleurs normalisées de pixels sont centrées autour de zéro; (2) ensuite, le fait d'appliquer une méthode linéaire comme l'ACP sur un ensemble des histogrammes rgb réduit la dimension de l'espace de représentation de ces données du 64 à 3, avec une qualité de représentation de 95%; dans ce nouvel espace la représentation des apparences d'objets n'est certainement plus assez discriminante et donc il y aura beaucoup de chevauchements entre les densités de mélange gaussien des différents modèles d'objets suivis, ce qui explique les classements incorrects des requêtes. Aussi, on peut rajouter à ce qui précède que la base de tests contient des changements de luminosité naturelle (soleil, ombrage, etc., voir figure 4.5) et qui ne correspondent généralement pas aux modèles théoriques implémentés ici. Également, les changements de luminosité sont appliqués partiellement sur les objets, ce qui rend critique l'application de la normalisation de couleurs. En effet, ce type de changements de luminosité devra peut être pris en compte par les chercheurs qui travaillent sur ce sujet de normalisation de la couleurs, qui sort du cadre de cette thèse. Enfin, l'invariance repose sur un modèle théorique qui supprime brutalement une information significative : sous prétexte de supprimer l'incidence d'une variabilité, somme toute limitée, de la luminance, on supprime complètement cette information.

Une solution pourrait être envisagée : considérer des histogrammes de couleurs rgb de dimension plus élevée (4096 classes de couleurs par exemple). Cependant, ceci est loin d'être validé dans nos expérimentations car le nombre des apparences est très limité et nos expérimentations montrent qu'une telle situation conduit à une estimation instable.

Enfin, rappelons que la segmentation d'objets (durant la phase de suivi) n'est pas fiable et elle contribue à la variabilité des objets et ensuite aux mauvais classements des requêtes.

Descripteur	d	$ci.\%$	$cr.\%$
RGB	64	61.4	59.3
HSV	64	59.0	60.0
HS	49	52.9	56.6
I	32	37.8	43.2
H	32	53.7	56.1
S	32	42.3	40.8
rgb	64	30.4	25.6
$l_1l_2l_3$	64	45.5	48.3

TAB. 4.4: *Résultats des scores de bon classement par la méthode de "moyenne temporelle de descripteurs" dans l'espace initial de descripteurs.*

Projection des données. Il est très difficile de conserver la même variance de données lors d'une projection dans un espace de petite dimension (section 3.4). L'analyse en composante principale ne prend pas en compte la structure non linéaire des données, des structures contenant des groupes ayant des formes arbitraires. Visuellement, il est impossible de vérifier si nos données ont des structures particulières dans les espaces à 64 dimensions, mais elles ne sont probablement pas linéaires. Ceci peut être validé par exemple par le fait que les pourcentages de bons classements obtenus dans l'espace initial de descripteurs avec la méthode de la "moyenne temporelle de descripteurs" et en utilisant la distance de χ^2 (formule 3.3) normalisée pour l'appariement sont plus élevés que dans l'espace réduit (voir tableaux 4.4 et 4.3).

Par contre, une méthode non-linéaire comme l'analyse en composante curvilignes (ACC) [30] et la méthode *multidimensionnel scaling* (MDS) [73] peuvent être employées. Cependant, des tests menés dans notre équipe ont montrés que l'ACC ne donne pas des meilleurs résultats que l'ACP [33]. D'autre part, la méthode MDS est coûteuse et n'est pas bien adaptée pour des données de taille importante (par exemple il nous a fallu plus de 3 heures pour une matrice de données de taille 1391×64 lorsque l'optimisation de la fonction de coût converge vers une solution).

4.7 Conclusion

Ce chapitre a présenté une approche de classification supervisée des objets suivis, où l'auteur de la vidéo hyperliée intervient dans la sélection des "modèles d'objets suivis" pour une séquence vidéo traitée. Le mélange gaussien est employé dans ce cas de discrimination. Mais son utilisation n'est pas classique ici, car chaque "modèle d'objet suivi" (une classe en discrimination) est lui même représenté par un mélange de modèles gaussiens : les classes d'apparences intra-plan modélisées par des gaussiennes. Rappelons que la structure de ces classes d'apparences intra-plan est déterminée automatiquement comme nous l'avons vu dans le chapitre précédent.

Après l'identification de ces classes d'apparences intra-plan de tous les modèles d'objets suivis, une loi globale de mélange gaussien est construite par un nouveau calcul des paramètres des proportions seulement. Le classement des nouvelles apparences d'objets dans la vidéo est effectué par deux méthodes : classement par maximum a posteriori et classement robuste par vote majoritaire.

Les expérimentations menées ici sur huit types de descripteurs globaux (histogrammes de couleurs bruts et normalisés) permettent de tirer les conclusions suivantes :

- La méthode de classement robuste donne de meilleurs résultats d'un ordre de 10% environ que celle du maximum a posteriori.
- La méthode de classement robuste donne des résultats bien meilleur d'un ordre allant jusqu'à 35% que la méthode classique de la "moyenne temporelle de descripteurs".
- Les résultats sont du même ordre quand les deux critères BIC et ICL sont utilisés pour le choix automatique de la structure des classes d'apparences intra-plan. Par contre lorsque le critère NEC est utilisé les résultats sont très mauvais. Ce critère n'est pas adapté au choix du modèle gaussien.
- Les résultats sont les meilleurs lorsque les classes d'apparences intra-plan sont recherchées dans l'espace des histogrammes de couleurs RGB réduit par ACP. La réduction de l'espace des histogrammes normalisés rgb conduit à une perte d'information significative ($d_E = 3$ avec une qualité de représentation des données de 95%) et ensuite à un chevauchement très large entre les classes d'apparences intra-plan des différents modèles d'objets suivis. Ceci explique les mauvais résultats de classement obtenus dans cet espace. Une telle réduction doit être évitée.

Une méthode alternative pour palier le problème de données limitées, et donc pour éviter la réduction de l'espace par ACP est de générer peut-être des apparences virtuelles des modèles d'objets suivis. La question qui se pose à ce stade : quel modèle mathématique faut-il appliquer pour générer des apparences virtuelles ? un réponse simple possible est la fabrication automatique d'occultation partielle. Une autre méthode qui nous paraît naturelle est d'effectuer l'apprentissage d'une classe d'objets sur plusieurs apparences d'un même objet suivi dans différents plans.