Chapitre 3

Modélisation statistique de l'apparence intra-plan des objets vidéo

Il faut savoir
perdre du temps pour en gagner.

Dans le chapitre précédent, nous avons présenté les outils de base que nous utilisons pour partitionner la vidéo en plans, pour localiser – automatiquement ou manuellement – des objets dans les images, et pour les suivre à travers le temps.

Ce chapitre présente principalement une étude du problème de la variation de l'apparence des objets suivis à travers le temps, et les conséquences de cette variation sur la reconnaissance des objets vidéo. Ensuite, il expose l'approche que nous avons retenue pour la modélisation statistique de cette variation dans l'espace de descripteurs. Les modèles statistiques issus de cette modélisation seront utilisés dans la suite de ce travail pour indexer, apparier, et classifier les objets suivis d'une séquence vidéo.

3.1 Introduction

Le contenu dynamique d'un film vidéo au sens d'animation et mobilité des objets à travers le temps (acteurs, voitures, etc.) représente le point clé qui attire l'attention des spectateurs d'une part et les analystes de la vidéo d'autre part. L'illusion d'animation des scènes est due à la projection à grande vitesse des images fixes (de 24 à 32 images par seconde). Pour rapprocher les actions du film à la réalité, les scènes sont réalisées dans des conditions naturelles d'éclairage et d'environnement. Aussi, les acteurs d'une scène sont filmés sous une grande variété de prise de vues, de zooms et de formes.

3.1.1 Les problèmes adressés

• **Apparence et variabilité intra-plan des objets suivis** Partons d'un film vidéo partitionnée au préalable en plans cinématographiques et en objets. Les objets sont suivis dans les plans. Du point de vue vision artificielle par ordinateur, les changements d'apparence des objets les plus fréquents dans un film vidéo peuvent être résumés comme :

- *Changement de point de vue de l'objet* suivi à travers le temps, à l'intérieur des images d'un plan vidéo et/ou dans différents plans. Un changement de point de vue de l'objet correspond à une rotation $3D$ (3 degrés de liberté) et/ou à une translation (3 degrés de liberté). Les rotations $2D$ des objets dans l'image (3 degrés de translations et 1 seul degré de rotation) sont rares dans la vidéo.

- *Changement d'échelle de l'objet* d'un plan à un autre sans connaissance a priori du degré de zoom.

- *Changement de la scène* ou *Occultation partielle* de l'objet non-rigide suivi à travers le temps. Des parties sont occultées et d'autres dévoilées durant le mouvement de l'objet suivi dans le plan ou durant son apparition dans différents plans du film.

- *Changement d'éclairage* des scènes de la vidéo (éclairage artificiel ou naturel, direct ou indirect). L'obscurité de la nuit et les ombres portées ne correspondent généralement à aucun modèle d'éclairage mathématique : surfaces et reflectances sont trop complexes et surtout inconnues.

- *Changement des conditions d'enregistrement* qui inclut le flou de l'image et bruits du signal. Généralement, ces changement ne sont ni contrôlables ni prévisibles.

En réalité les objets mobiles d'un film vidéo sont contaminés par une combinaison de ces différents types de changements d'images. La figure 3.1 illustre quelques apparences intra- et inter-plan d'un objet non-rigide et en mouvement (la voiture blanche entourée). Ces images sont extraites du corpus vidéo *Avengers* (section 2.5).

Sous ces conditions difficiles, mais réelles, de changements de l'apparence des objets, il est évident que la reconnaissance des objets similaires inter-plan est une tâche très délicate. Aussi à l'intérieur d'un plan vidéo il est souvent difficile d'identifier visuellement les occurrences d'un même objet suivi. Par exemple la première ligne de la figure 3.1 met en évidence les grandes variations de point de vue et la grande variété d'occultations qui peuvent survenir.

La représentation spatiale d'un objet suivi dans l'espace de descripteurs est directement influencée par ces changements d'apparences intra-plan. Supposons que chaque occurrence de l'objet suivi est représentée par un seul point (vecteur de descripteurs) dans un espace de descripteurs multidimensionnels. Aux changements d'apparence de l'objet d'un instant à un autre correspondent une trajectoire ou un nuage de points dispersés dans l'espace. Dans la suite cet phénomène est appelé "**variabilité de l'objet suivi dans l'espace de descripteurs**". Selon la progression du changement d'apparence de l'objet à travers le temps une certaine continuité de la trajectoire peut être conservée ou non.

FIG. 3.1: *Exemple d'apparences de la voiture suivie dans la séquence vidéo Avengers*

La figure 3.2 illustre la variabilité intra-plan d'un objet mobile dans l'espace de couleurs. Dans cette séquence de publicité "Ajax" la caméra est fixe. L'objet d'intérêt est mobile : l'enfant court et se déplace vers la caméra. Donc au fur et à mesure qu'il se rapproche d'elle une partie de son corps est occultée pour qu'à la dernière image du plan, seule sa tête soit visible. Aussi, l'éclairage du soleil est bien présent sur les images du début du plan et totalement absent sur les images de la fin. La figure 3.2.a représente les quatre vues de l'enfant aux instants $1, 10, 16$ et 25 parmi 26 vues en total dans le plan. La figure 3.2.b représente la distribution des couleurs de chacune de ces quatre vues (chaque pixel est représenté par ces trois valeurs de couleurs : rouge, verte et bleu). Sans aucune idée sur les images de l'enfant, cette dernière figure permet à un analyste de conclure que ces distributions correspondent au même objet image, mais sous différentes conditions de changements d'images. D'autre part, sur la figure 3.2.c on remarque que les couleurs dominantes sont placées dans des positions différentes. Cela explique les occultations partielles et/ou les changements d'éclairages que l'enfant a subit à travers le temps. Une couleur dominante peut être vue comme étant la couleur ayant la fréquence la plus élevée. Chaque représentation $3D$ de la figure 3.2.c correspond à un histogramme calculé dans l'espace de couleur RGB de l'enfant a l'instant t ($t = 1, 10, 16, 25$). Chaque couleur (R, G, B) est représentée par un carré dont la taille est proportionnelle à la valeur de sa fréquence. La figure 3.2.d illustre cette variabilité sur la seule dimension du premier axe principal de l'histogramme RGB par rapport au temps (instants [1..26]). Visuellement, il est clair

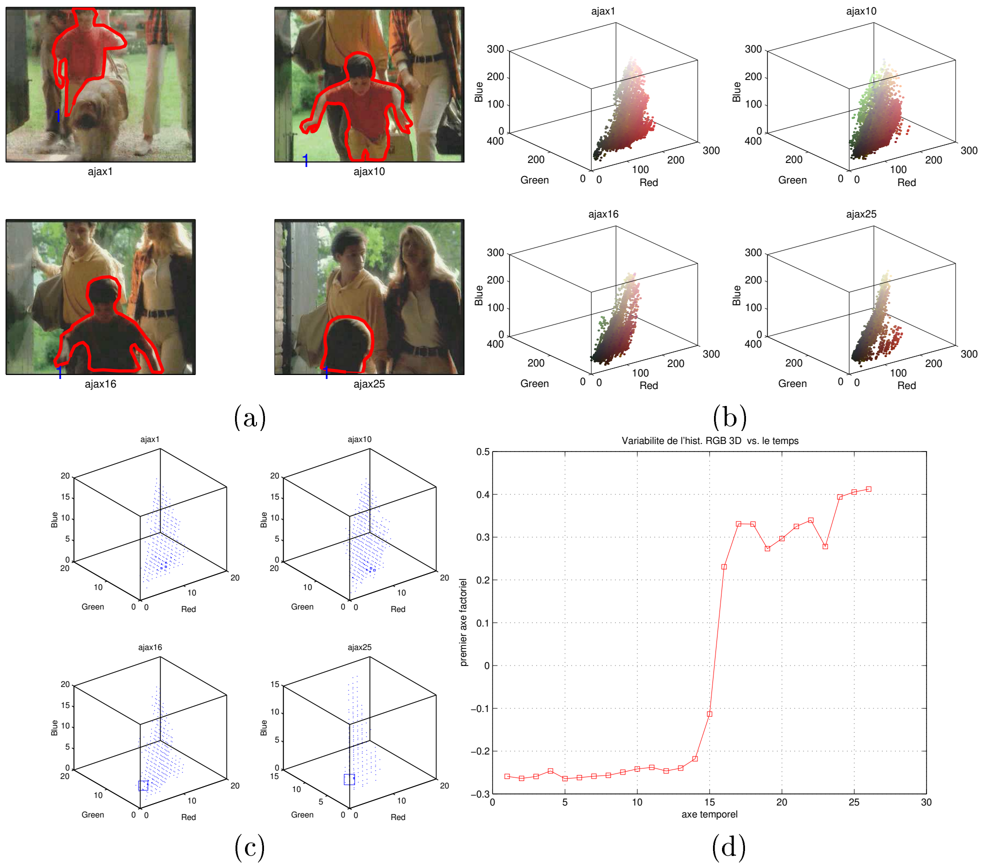


FIG. 3.2: *Exemple de la variabilité intra-plan de l'objet suivi dans l'espace de descripteurs : (**a**) Quatre vues de l'apparence de l'enfant d'Ajax aux instants 1, 10, 16 et 25 (26 occurrences en total); (**b**) Les histogrammes RGB correspondants; (**c**) Les histogrammes RGB correspondants où chaque couleur (R, G, B) est représentée par un carré dont sa taille est proportionnelle a la valeur de sa fréquence; (**d**) Illustration de la variation du premier axe principal des histogrammes par rapport au temps.*

que la distribution de l'objet dans cet espace unidimensionnel est probablement bimodale (cette distribution peut être approximée par deux segments de droites parallèles).

• **Taille de la vidéo et indexation des objets suivis** Selon les normes, 24 à 32 images sont projetées par seconde. Si on considère que dans chaque image un unique objet est localisé, un film vidéo de 90 minutes comporte environ 129600 objets à indexer. Ce problème d'indexation devient ingérable en terme de stockage physique et de temps de recherche lorsque les descripteurs extraits des ces objets sont de grandes dimensions. Par exemple, Schiele dans sa thèse [108] utilise des histogrammes multidimensionnels à champs réceptifs de 10^9 cellules (histogramme à six axes et d'une résolution de 32 cellules par axe) et chaque cellule est codée sur 64 bits.

Les solutions proposées dans la littérature pour palier à ce problème sont basées principalement sur la réduction du nombre des images de chaque plan en représentant l'ensemble de ces images par des images clés [92]. Le choix des images clés les plus représentatives du contenu du plan est un problème qui sera abordé dans le chapitre 6 de ce mémoire [58]. Souvent l'image médiane est considérée comme image clé du plan. Selon cette approche, seule l'occurrence médiane d'un objet suivi dans le plan est indexée. En conséquence, la taille de la base d'objets à indexer est divisée par un facteur 129 si on suppose qu'en moyenne un film vidéo est divisé en 1000 plans. Généralement, le nombre de plans d'un film vidéo varie entre 500 et 1000 plans [70] [1]. Notons que pour les films d'action le nombre de plan est plus élevé car la durée des plans est très courte (environ une seconde). Par exemple, le film *Octobre* de S.M. Eisenstein est décomposé d'environ 3225 plans [68].

Cette solution d'images clés peut être envisageable pour le problème d'indexation des objets vidéo posé ici si ces objets sont statiques et si leurs variations intra-plan dans l'espace de descripteurs sont négligeables. Mais malheureusement la plupart des objets d'intérêts des films vidéo sont mobiles et leur variabilité intra-plan dans l'espace de descripteurs est significative (voir figure 3.2). Une étude expérimentale menée sur une séquence vidéo réelle (qui sera détaillé dans le chapitre 4) a montré l'inefficacité de cette méthode pour la reconnaissance d'objets vidéo [55] [57].

3.1.2 La solution proposée

Afin de résoudre les deux problèmes exposés ci-dessus, nous proposons de *modéliser* la variabilité d'un objet suivi dans l'espace de descripteurs de couleurs par une fonction statistique de *mélange de lois*, qui **capture** les différentes apparences intra-plan de l'objet. L'idée de base est de classifier les occurrences semblables en terme d'apparence (dans l'espace de descripteurs) en des groupes. Ensuite chaque groupe construit sera représenté par ses *paramètres statistiques*, ce qui **réduit** considérablement le nombre des descripteurs à indexer par objet suivi. Rappelons que chaque occurrence d'un objet suivi est supposé être représentée par un seul point multidimensionnels dans l'espace de descripteurs. La seule hypothèse que nous fixons dans la suite de ce travail est que la forme de la distribution dans l'espace de descripteurs est capturée par un mélange de gaussiennes, hypothèse que nous discuterons plus tard. Les changements d'apparence de l'objet suivi seront alors modélisés par une *densité de mélange gaussien* dont le nombre des composantes dépend de la complexité de ces changements.

3.1.3 Organisation du chapitre

Dans une première partie nous discuterons les différents descripteurs que nous pourrons choisir pour décrire les apparences intra-plan des objets suivis. Puis notre approche de modélisation de la variabilité intra-plan sera présentée en détail. Dans la section 3.7 quelques résultats expérimentaux d'objets suivis modélisés sont illustrés. Les commentaires et les conclusions d'une telle approche sont discutés dans les deux dernières sections de ce chapitre.

3.2 Représentation de bas niveau des objets : un état de l'art

Dans le cadre de ce travail on se limite à la définition suivante d'un descripteur (définition 3.2.1), mais selon le prochain norme MPEG-7 un descripteur peut avoir d'autres formes, par exemple l'URL de l'image, le texte détecté dans les images d'un film vidéo, etc.

Définition 3.2.1 *Un descripteur est un vecteur de réels, mono- ou multidimensionnels, qui résume d'une manière efficace une ou plusieurs caractéristiques du contenu de l'image (comme par exemple, la couleur, la forme et/ou la texture).*

L'extraction des descripteurs de bas niveau, des images ou des régions d'images, est l'étape de base pour tout système de recherche d'images par le contenu (CBIR [1]). Un descripteur est associé à chaque image (objet) indexé dans la base d'images (d'objets). La recherche d'une image requête consiste premièrement à extraire un descripteur de cette image et ensuite de le comparer avec ceux de la base. Les images de la base les plus proches de l'image requête, en terme de distance de similarité, sont sélectionnées, présentées et ensuite elles peuvent être parcourues par l'utilisateur final du système CBIR.

Dans la suite, quelques propriétés fondamentales des descripteurs sont rappelées. Puis une courte présentation de l'état de l'art des catégories de descripteurs intégrés dans des systèmes de recherche d'images et des systèmes de vidéothèque est faite.

3.2.1 Propriétés d'un descripteur

Dans le domaine de l'indexation d'une base d'images fixes ou d'objets, quelques propriétés de descripteurs sont données pour pouvoir étudier leurs performances en pratiques [94] [63]. Essentiellement, un descripteur doit posséder les trois propriétés suivantes :

1^0) *Discriminant* pour mieux identifier les images similaires et rejeter les images différentes. Si *f(I)* est le descripteur de l'image I, $\mid f(I) - f(I') \mid$ doit être suffisamment grand dès que I et I' ne sont pas similaires (ici $\mid . \mid$ est une distance).

2^0) *Complexité faible* en terme de temps de calcul. Le calcul de $f(.)$ doit être rapide.

3^0) *Taille raisonnable* pour gérer facilement une large base d'images et pour accélérer la procédure de comparaison des descripteurs. La taille du descripteur et le temps de comparaison sont deux facteurs proportionnels.

Une image (objet) indexée et/ou recherchée dans une base d'images (objets) peut être présente sous une variété de prises de vues, de changements d'éclairages, de rotations et d'occultation partielles de son contenu. Sous ces conditions de changements la puissance de discrimination d'un descripteur est fortement liée à son degré d'*invariance* ou de *robustesse*.

1. Content Based Image (and Video) Retrieval

Un descripteur est dit *invariant* à tel type de changement d'image lorsque sa valeur est la même avant et après ce changement : $f(I) = f(t(I))$ où t représente la transformation que l'image I a subite (plus d'informations sur la théorie d'invariance dans [86]). Dans le cas ou la valeur du descripteur a peu varié après ce changement, le descripteur est dit *robuste*; cette robustesse est surtout essentielle en cas de bruit non modélisable comme l'occultation partielle.

3.2.2 Descripteurs pour l'appariement d'images

La première différence entre les systèmes traditionnels de base de données et les base d'images (et de vidéo) est la façon avec laquelle les requêtes sont exprimées. Tandis que les langages d'interrogation standards comme SQL (*structured query language*) peuvent efficacement exprimer l'intention de l'utilisateur, il est très difficile qu'un utilisateur exprime le contenu d'une image. Les requêtes tendent à être exprimées dans des langages naturels et sont elles-mêmes tout à fait complexes. Face à ce problème, la communauté de vision par ordinateur et de reconnaissance de formes a développé des méthodes pour convertir la spécification de requête en un jeu significatif de descripteurs de requête (*Query features*), qui peuvent être appliqués pour rechercher les données visuelles.

Les utilisateurs d'un tel système ne peuvent pas comprendre les méthodes complexes employées pour mettre en application la recherche d'images. Par conséquent une interface plus intuitive doit être donnée aux utilisateurs, de telle sorte que la requête puisse alors être traduite en des paramètres appropriés du système de recherche d'images.

Tout ce qui précède mène au besoin d'une couche intermédiaire entre la requête utilisateur et le processus de recherche qui transmet la requête par l'utilisation d'un descripteur approprié pour interroger l'image. Ces descripteurs devraient être calculés en temps réel et également être capables de capturer l'essence de la requête humaine.

Plusieurs systèmes multimédia développés ces dernières années sont décrits dans la littérature [48] [3]. Parmi ces derniers, les systèmes performants incluent le projet QBIC (*Query By Image Content*) [91], le système Photobook [94], le système MMIS (*Manchester Multimedia Information System*) [47], VisualGREP [76], le projet IDL (*Informedia Digital Library*) [115], SurfImage [90] et celui développé dans notre laboratoire [110].

Par exemple, le système QBIC développé par le centre de recherche *Almaden* de *IBM* utilise une variété de descripteurs pour rechercher les images dans une bases d'images fixes [91]. Il permet à l'utilisateur de rechercher une image requête, que ce soit une image appartenant à la base d'images ou une image d'extérieur ou dessiner manuellement la forme recherchée (*Query-by-Example*), de visualiser et de parcourir (*browse & navigate*) les résultats de la requête (un ensemble d'images). Des descripteurs visuels telles la couleur et la texture sont précalculés et indexés pour toutes les images de la base. Une extension de ce système pour interroger une base vidéo à récemment été rajoutée.

Le système VideoPrep développé dans le cadre de ce travail sur un contrat entre l'INRIA (projet MOVI de l'unité Inria Rhône-Alpes et VISTA de l'unité Inria Rennes) et le centre de recherche d'Alcatel CRC, et destiné à la création de la vidéo hyperliée, utilise des descripteurs globaux de la couleur et des descripteurs locaux de niveaux de gris. Il permet à l'utilisateur de sélectionner un objet d'intérêt dans le film vidéo, de chercher toutes ces apparitions dans le film, de naviguer et de visualiser les résultats (plus de détails dans le chapitre 8).

Les méthodes d'appariement d'images et d'objets décrites dans la littérature utilisent trois types de descripteurs extraits à partir de l'image. Ces descripteurs sont fondés principalement sur la couleur, la forme et la texture. Quelques systèmes d'indexation, comme décrits ci-dessus, emploient un ou plusieurs descripteurs pour l'appariement des images.

3.2.2.1 Descripteurs globaux de couleur

La couleur est la caractéristique la plus exploitée par les systèmes d'indexation et de recherche d'images par le contenu. Swain et Ballard [120] ont proposé d'utiliser l'histogramme pour décrire la distribution globale des couleurs d'une image.

Un histogramme est un outil statistique qui code une population sous la forme des fréquences (ou effectifs) des individus. Lorsque une image est de haute résolution (16.7 millions de couleurs) une quantification de couleurs est généralement effectuée afin de réduire le nombre de cellules (classes de couleurs) de l'histogramme. La méthode de quantification la plus simple consiste à fusionner les classes de couleurs situées dans un intervalle de couleur donnée. Une autre consiste à appliquer un algorithme de clustering pour fabriquer les classes de couleurs d'une image. Le choix du nombre de cellules est un point critique pour l'appariement d'images par la méthode d' histogramme. Les expérimentations menées dans [108] et [45] montrent que les histogrammes de dimensions réduites fournissent des bon résultats d'appariements.

Les motivations principales d'utilisation de l'histogramme sont la rapidité de son calcul et son indépendance à plusieurs changements d'images : position relative des objets dans l'image, rotation 2D des objets et le changement d'échelle. Également, le calcul de l'histogramme dans un espace de couleurs normalisées, comme par exemple l'espace de chrominances (rgb), le rend invariant aux changements d'éclairages [116] [36]. La transformation de l'espace de couleur RGB en rgb normalisé est donnée par :

$$\begin{cases} r(R,G,B) = \frac{R}{R+G+B} \\ g(R,G,B) = \frac{G}{R+G+B} \\ b(R,G,B) = \frac{B}{R+G+B} \end{cases} \tag{3.1}$$

Cette normalisation de la longueur de chaque couleur RGB (donne $R+G+B=1$) est une méthode performante pour limiter la dépendance de l'intensité [36].

Pour apparier une image requête avec celles de la base d'images, Swain et Ballard calculent les histogrammes dans l'espace de couleurs HSV (équation 3.2), puis utilisent l'intersection d'histogrammes pour les comparer. On a :

$$\begin{cases} H(R,G,B) = \arctan\left(\sqrt{\frac{\sqrt{3}(G-B)}{(R-G)+(R-B)}}\right) \\ S(R,G,B) = 1 - \frac{min(R,G,B)}{I(R,G,B)} \\ V(R,G,B) = \frac{R+G+B}{3} \end{cases} \tag{3.2}$$

où H représente la teinte qui est la caractéristique la plus importante de la couleur, S représente la saturation qui mesure le contenu relatif du blanc d'un couleur (i.e. la saturation du blanc est égale à zéro, la saturation d'une couleur pure est égale à l'unité et la saturation du noir est indéfinie), et V représente la luminance ou l'intensité définie par la valeur portée par l'axe achromatique du cylindre de MUNSELL.

Notons qu'il n'y a pas des critères théoriques permettant de sélectionner la distance la plus appropriée pour comparer des histogrammes. Schiele dans sa thèse ([108], chapitre 4) a comparé plusieurs distances. Il conclut, sur une base expérimentale d'objets rigides, que la distance du χ^2 normalisée fournie les meilleurs résultats d'appariements :

$$\chi^2(Q,V) = \sum_i \frac{(q_i - v_i)^2}{(q_i + v_i)} \tag{3.3}$$

où q_i et v_i représentent les fréquences de la i^{eme} cellule des histogrammes Q et V respectivement.

Plusieurs améliorations ont été apportées aux histogrammes, soit pour qu'ils soient moins sensibles aux changements d'éclairage, soit pour qu'ils incorporent une certaine information géométrique. Funt et Finlayson [39] ont proposé d'utiliser les dérivées de logarithmes des canaux afin de fournir des caractéristiques invariantes aux changements d'éclairage. Leur objectif est de limiter les effets de la luminosité en normalisant les images vers une illumination standard. Healey et Slater [113] ont aussi proposé des algorithmes du même genre. Ils calculent des invariants de moment de l'histogramme de couleurs en entier. Ces invariants sont fondés sur le modèle linéaire de dimension finie de couleurs permettant la modélisation des changements de l'intensité d'éclairage par une transformation linéaire d'histogrammes.

D'un autre coté, plusieurs auteurs ont proposé d'incorporer des informations spatiales avec la couleur [117] [101] [93]. La plupart des méthodes proposées divisent l'image en régions. Une récente méthode nommée CCV (color coherent vector) [93] utilise une autre approche (raffinement de l'histogramme) : les cellules de l'histogramme sont partitionnées en fonction de la cohérence spatiale des pixels; un pixel étant cohérent s'il appartient à une région d'assez grande taille et uniformément colorée, il est incohérent sinon. Un CCV représente une classification pour chaque couleur dans l'image. Il est facilement implémentable et donne de bien meilleurs résultats que les histogrammes. Par contre, il est sensible aux occultations partielles et aux changements d'illuminance.

Par ailleurs, une autre méthode, basée ni sur le partitionnement de l'image en régions, ni sur un raffinement des histogrammes, et qui semble donner de très bons résultats, a été proposée : il s'agit des corrélogrammes [63]. A cheval entre les méthodes purement locales comme les points d'intérêts et celles purement globales comme la distribution des couleurs, les corrélogrammes tiennent compte aussi bien de la corrélation spatiale des couleurs que de la distribution globale de cette corrélation. Les corrélogrammes apparaissent ainsi comme une solution efficace pour la recherche d'images par le contenu dans une grande base d'images. Selon les concepteurs de cette méthode, une évidence expérimentale suggère que cette nouvelle signature dépasse non seulement les performances de la méthode des histogrammes de couleurs mais aussi les méthodes récentes proposées pour l'affiner et l'améliorer dans le cadre de l'indexation et de la recherche d'images. Un corrélogramme de couleur décrit comment la distribution de probabilité conditionnelle de paires de couleurs change avec la distance. Il est défini comme étant une table indexée par paires de couleurs telle que la k^{eme} entrée pour (i, j) représente la probabilité de trouver un pixel de couleur j à une distance k à partir d'un pixel de couleur i dans l'image. La méthode par corrélogrammes incorpore d'une part la corrélation spatiale des couleurs et décrit d'autre part la distribution globale de la corrélation spatiale locale des couleurs.

Par contre, les corrélogrammes ont des tailles importantes (le minimum conseillé est de 320 pour une distance spatiale d qui vaut 5 et une palette de couleurs m de $4 \times 4 \times 4$). Ils sont aussi sensibles aux occultations partielles et aux variations spatiales du contenu de la scène (déplacement des objets d'une image). D'autre part, le calcul d'un corrélogramme est très lent à cause du balayage de l'image pour les distances spatiales de 1 à d. Les auteurs de ce descripteur proposent d'utiliser la programmation dynamique pour accélérer son calcul. Ils proposent également l'autocorrélogramme qui est comme le corrélogramme dans lequel on ne considère que des paires de couleurs identiques. Son temps de calcul est relativement raisonnable. Par contre, il est moins informatif et donc discriminant que le corrélogramme.

Un autre type de recherche propose de ne pas se baser uniquement sur la couleur mais de faire des histogrammes d'autres caractéristiques locales. Construire un histogramme des résultats de filtres spatiaux orientés ou non est un exemple. Bernt Schiele a proposé une étude qui semble très intéressante à ce sujet [108]: il construit un histogramme multidimensionnels de champs réceptifs et sa méthode semble donner de bons résultats. Cependant ce descripteur a une dimension trop grande et cela représente un handicap majeur pour leur utilisation dans un système d'indexation d'une large base d'images.

Plus récemment encore, une étude de Gevers et Smeulders, qui semble prometteuse, propose une manière efficace de coupler la distribution des couleurs et les invariants géométriques [116] [46]. Dans leur système PicToSeek, les auteurs utilisent différents modèles d'illuminations, la saturation de la couleur, la transition de la couleur, le fond de l'image, un invariant géométrique basé sur le birapport et des histogrammes autour des points d'intérêts détectés dans l'image contour. La transition de la couleur se rapporte au nombre de changements de tonalité (hue) de l'image. Une image avec beaucoup de détails aura plus de transitions qu'une image avec peu de changements. Le fond est supposé être celui de l'image. C'est la cellule de l'histogramme ayant la valeur la plus élevée. L'espace de couleur HSV est choisi car il contient le paramètre V (intensité) représentant l'intensité dans l'image, et pouvant être étendu à la luminosité dans la scène. Un modèle de réflexion pour l'objet est établi et des invariants photométriques de la couleurs sont calculés. Comme nous le verrons dans les expérimentations (chapitres suivants), le modèle $l_1 l_2 l_3$ défini par l'équation 3.4 ([43]) est adopté dans la génération des descripteurs de test.

$$\begin{cases} l_1(R,G,B)=\frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \\ l_2(R,G,B)=\frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \\ l_3(R,G,B)=\frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \end{cases} \tag{3.4}$$

Le système PicToSeek est décrit en grand détail dans [44]. La base de test utilisée est fabriquée par les auteurs de ce système. Les objets sont acquis par un appareil photo de haute qualité et dans des conditions d'éclairage bien choisies. Il paraît cependant que les résultats d'appariements varient largement d'un descripteur à un autre, ils sont mauvais lorsque des rotations des objets 3D dans les images requêtes sont présentes.

3.2.2.2 Descripteurs de la forme

La caractéristique la plus importante pour la forme est son contour (2D) ou son enveloppe convexe (3D). Différentes approches d'appariement de la forme sont adoptées par les systèmes de recherche d'images par le contenu. Jain et Vailaya [67] proposent d'apparier la forme sur la base d'un histogramme de direction du gradient aux points de contour de l'image. Le filtre de Canny [19] est appliqué pour obtenir l'image contour. Les histogrammes sont comparés en utilisant l'intersection d'histogramme comme distance. Un histogramme de direction de gradient est invariant à la translation. Il est également invariant aux changements d'échelles s'il est normalisé par le nombre de points du contour. Cette méthode est rapide mais trop sensible aux bruits du signal, aux occultations partielles et aux changements d'éclairages qui affectent beaucoup la détection du contour. Rui et al. [106] utilisent les descripteurs de Fourier pour décrire la forme. Par contre, la transforme de Fourier discrète n'est pas invariante aux transformations affines. Les auteurs définissent un descripteur modifié de Fourier qui est une forme interpolée des coefficients de basse fréquence de descripteurs de Fourier normalisés. L'appariement des images s'effectues en utilisant plusieurs distances comme la distance euclidienne et la distance de Hausdorff. D'autre méthodes basées sur une approximation du contour par des fonctions mathématiques précises tels les B-Splines, sont aussi employées [37] [26]. Pour plus de détails sur la caractérisation de la forme, les lecteurs peuvent se reporter à l'article de Loncaric [77].

3.2.2.3 Descripteurs de la texture

Les caractéristiques visuelles des régions homogènes des images réelles sont souvent identifiées comme texture. Les moments du deuxième ordre, l'énergie, l'entropie, la corrélation, l'homogénéité locale, l'inertie et le contraste, dérivés de la matrice de co-occurrences de niveaux de gris, sont les plus connus pour décrire la texture. Une matrice de co-occurrences est un histogramme à quatre dimensions : $S = f(i, j, d, \theta)$ où i et j sont les niveaux de gris des pixels à une distance d, et θ est l'angle formé par la ligne qui rejoint les deux pixels et l'axe horizontal. Cependant, le calcul de la matrice de co-occurrences est très coûteux surtout lorsque la distance spatiale d est relativement élevée. Notons que le corrélogramme est une extension de la matrice de co-occurrences pour les images couleurs. Lorsque l'orientation est prise en compte dans le calcul du corrélogramme, ce dernier est nommé “corrélogramme d'orientation”. Connors et Harlow [27] ont montré sur des exemples réels que ces six mesures de la texture ne sont pas suffisamment discriminantes ([88] chapitre 5). Leur performance de discrimination augmente quand plusieurs valeurs de d sont utilisées. D'autre descripteurs comme les coefficients produits par le filtre de Gabor et les Ondelettes sont aussi utilisés pour décrire la texture [78]. Ils supposent que la texture des régions soit localement homogène. Cependant ces descripteurs ne sont pas invariants à la plupart des changements d'images. Notons que le système Photobook dû à

Pentland et al. [94] est quasiment le seul système de recherche d'image par le contenu qui est basé sur la texture. Les auteurs caractérisent une texture, à l'aide de la décomposition de Wold, par trois mesures de la périodicité, de la directionnalité et de l'aspect aléatoire (*Randomness*).

3.3 Choix de la base de descripteurs : nos données

Les descripteurs globaux discutés ci-dessus sont nullement complets. D'autres catégories de descripteurs existent : les descripteurs locaux par exemple. Les images sont caractérisées par des descripteurs locaux du signal RGB, tels qu'ils ont été définis par Gros et al. [49], lorsqu'ils ont étendus le travail de Schmid et Mohr [110]. Ceux-ci sont des vecteurs de dimension 24 et ils représentent des mesures locales invariantes aux transformations géométriques (rotation, translation, changement d'échelle) et aux changements de la luminosité. Ces mesures sont calculées autour des “point d'intérêt” obtenus par le détecteur de Harris [59].

L'étude de la variabilité présentée ici n'est pas étendue pour les descripteurs locaux mentionés ci-dessus. Une apparence d'objet suivi est représentée par plusieurs vecteurs descripteurs locaux calculés autour des points d'intérêts, entre deux apparences successives la mise en correspondance entre les points d'intérêts est une tâche difficile (objet mobile) et demande un suivi des points d'intérêts pour pouvoir étudier leur variabilité. De plus, ces descripteurs locaux sont bons pour la mise en correspondance mais pas pour l'indexation d'objets en mouvement complexe : des expérimentations que nous avons menées sur une base d'objets vidéo ont montré une performance très faible de ces descripteurs [18].

Les descripteurs globaux proposés jusqu'à ce jour manquent encore de robustesse aux différents changements d'images, et en particulier aux bruits et aux occultations partielles. Leur performance se dégradent rapidement lorsque il s'agit d'apparier des objets réels en mouvement en utilisant une méthode d'appariement classique (voir section 4.5). Ceci est du principalement à l'acquisition des objets non-rigides dans un environnement bruité et à leur apparences variables à travers le temps (voir figure 3.1).

Les insuffisances que nous venons de signaler expliquent que l'extraction de descripteurs reste un domaine de recherche très actif. Aussi, lorsqu'il s'agit de valider une nouvelle approche de reconnaissance ou d'indexation d'images, le choix de la base de descripteurs paraît comme une occupation prioritaire de certains chercheurs.

L'approche retenue dans ce chapitre pour capturer la variabilité intra-plan des objets suivis ainsi que nos approches de classifications de ces objets sont adaptées et expérimentées uniquement pour une base de descripteurs globaux.

Le descripteur global le plus utilisé est l'histogramme de couleur [120]. Rappelons que l'histogramme est invariant aux changements l'échelles, de translation, de l'orientation de la scène et de rotation des objets dans l'image. Aussi, après un certain nombre de traitements élémentaires de normalisation de couleurs (voir section 3.2) l'histogramme est rendu robuste à quelques types de changements d'éclairages. Par contre l'histogramme de couleurs est sensible aux occultations partielles, de même que tout les descripteurs globaux, et peut-être un peu moins que les descripteurs qui intègrent une information spatiale (les corrélogrammes par exemple). Un défaut de l'histogramme est que deux images différentes ayant la même distribution de la couleurs auraient des histogrammes similaires, et donc les deux images sont appariées semblables. Cependant, nous considérons que cette hypothèse à une probabilité très faible.

Malgré les défauts de l'histogramme, nous nous sommes contentés de le prendre comme un descripteur de base pour valider nos approches. Mais d'autres descripteurs tels que les corrélogrammes peuvent être envisagés.

La base de descripteurs utilisée est composée principalement des histogrammes calculés dans différents espaces de couleurs et de niveaux de gris. Ces espaces sont : RGB, HSV, H, S, I, rgb et $l_1l_2l_3$. Rappelons que les espaces rgb, H et $l_1l_2l_3$ sont invariants aux changements de luminance. Le choix de cette base est motivé par les propriétés intéressantes des histogrammes (voir section précédente), la nature non-rigide des objets exploités (difficulté d'adapter par exemple des descripteurs géométriques) et comme nous le verrons dans la suite le besoin des descripteurs de dimension réduite pour être capable d'estimer d'une manière stable la loi de mélange lorsque le nombre d'apparences intra-plan d'un objet suivi est limité. La réduction de la dimension des descripteurs est réalisée par une analyse en composante principale (section suivante).

3.4 Réduction de la complexité des données

Dans ce travail, et comme nous le verrons dans la suite, la dimension des descripteurs est un paramètre important à gérer. Lorsque un objet suivi possède un petit nombre d'apparences, la modélisation de sa variabilité dans un espace de descripteurs de grande dimension devient instable. Une réduction de la dimension de l'espace par une analyse en composantes principales (ACP) est un pré-traitement que nous appliquerons sur les données : tous les vecteurs descripteurs d'une base d'objets vidéo.

L'ACP est une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques, et qu'on utilise sans référence à des hypothèses de nature statistique ni à un modèle particulier [75]. Le but est d'étudier les modes de variations les plus importants, et qui caractérisent au mieux un nuage de points.

Cette technique est souvent utilisée par les approches de classification de visages [121]; une ACP est appliquée sur chaque classe de visages, un visage requête est projeté dans tous ces espaces et ensuite affecté à la classe la plus proche en terme de distance. La recherche est ainsi plus rapide que dans les espaces initiaux (moins de comparaison entre deux vecteurs descripteurs).

Principe Soit un tableau de données $Y = (y_{ij})$ formé de n lignes et de d colonnes. Les lignes représentent les individus (descripteur d'une apparence d'objet) et les colonnes représentent les variables (dimension d'un vecteur descripteur). Les lignes peuvent être considérées comme des réalisations indépendantes de vecteurs aléatoires de l'espace $\mathbb{R}^d$. La distance entre 2 individus a une interprétation géométrique directe. Il s'agit ici de la distance euclidienne classique entre deux points de $\mathbb{R}^d$. Deux individus sont proches (ou voisins) si et seulement si leurs d coordonnées sont proches. Deux individus proches ont un même comportement vis-à-vis des variables considérées, et on peut les mettre dans une même groupe. Considérons maintenant l'espace $\mathbb{R}^n$ des d variables. Si toutes les n coordonnées de deux variables sont proches, les variables seront représentées par deux points voisins dans $\mathbb{R}^n$. Cela veut dire que ces variables mesurent une même chose ou encore, qu'elles sont corrélées.

Analyse du nuage des individus Globalement, le nuage de points est approché par un hyper ellipsoïde. Le critère algébrique utilisé est que la somme des carrés des distances entre tous les couples d'individus soit maximale.

Les axes de l'ellipsoïde sont les directions des composantes principales et les longueurs des diamètres caractérisent la dispersion du nuage sur chaque composante. On les appelle valeurs propres de la matrice de covariance du nuage, qui, pour le tableau normalisé (centré-réduit), est donnée par : $C = (c_{jk})$ où $c_{jk} = \sum_{i=1}^{n} y_{ij}y_{ik} = cor(j,k) \leq 1$ représente le coefficient de corrélation empirique entre les variables j et k.

Les valeurs propres λ_j sont les termes situés sur la diagonale principale de la matrice obtenue par la diagonalisation de la matrice de corrélation : $D = (d_{jk})_{j,k=1..p}$ $d_{ij} = \lambda_j$ et $d_{jk} = 0$ si $j \neq k$. Les deux matrices sont liées par la relation : $C = P^T DP$ où P est la matrice dont les colonnes sont les vecteurs propres de la matrice de corrélations : $P = (u_1, \ldots, u_p)$. La projection des individus sur le sous espace factoriel E_k déterminé par k vecteurs propres liés aux plus grandes valeurs propres (composantes principales) forme un nuage de points, situé à l'intérieur de l'ellipse d'intersection entre le sous-espace E_k et l'hyper ellipsoïde.

Test de validité La qualité globale de représentation des données initiales sur le sous-espace factoriel E_k est mesurée par le pourcentage d'inertie (la variance) pris en compte par E_k :

$$Q_{E_k} = \frac{\displaystyle\sum_{j \in E_k} \lambda_j}{\displaystyle\sum_{i=1}^{d} \lambda_i} \times 100. \tag{3.5}$$

Les inerties associées aux λ_j peuvent être aussi représentées sur un histogramme. Si la décroissance est régulière, alors les données sont peu structurées et les variables faiblement corrélées, voir linéairement indépendantes. L'analyse ne fournira pas des résultats intéressants. Si, au contraire, la décroissance est assez irrégulière, présentant des paliers, visibles surtout sur les premières valeurs, alors il y a des fortes corrélations entre les variables et on peut réduire de beaucoup le nombre de composantes significatives. Un autre critère empirique nous permet de s'arrêter là où on trouve un coude sur l'histogramme.

Défauts de l'ACP La projection de données se fait d'une manière linéaire. Si ces données ont des structures complexes le risque de confusion augmente. En plus, le critère des moindres carrés qui est à la base de cette analyse suppose implicitement que les données ont une distribution gaussienne et si ce n'est pas le cas, les valeurs aberrantes ont une contribution significative à l'inertie des axes ce qui fausse les résultats de l'analyse. Un recours à des méthodes de pré-traitement des données pour éliminer les valeurs aberrantes est très utile (analyse de rangs par exemple). D'autres méthodes d'analyse de données non-linéaires comme par exemple l'analyse en composante curvilignes (ACC) [30] et la méthode *multidimensionnel scaling* (MDS) [73] peuvent être employées. Druga [33] dans son travail de DEA à l'équipe MOVI, a montré que l'ACC ne donne pas meilleurs résultats que l'ACP. Elle explique ceci par l'inexistence d'une forte structure dans les données (descripteurs d'invariants différentielles [49] [109]) et/ou par des relations linéaires ou linéarisables entre les composantes. Par contre, la méthode MDS est beaucoup moins utilisée car elle est trop coûteuse en terme de calcul.

3.5 Variabilité intra-plan des objets suivis : le problème

Avant de proposer une stratégie de classification des objets suivis, qu'elle soit supervisée ou automatique – objet des chapitres suivants –, il s'avère très utile de proposer une technique de caractérisation de bas-niveau, bien adaptée à nos données visuelles : les objets mobiles. La suite de ce chapitre se focalise sur ce point fondamental de notre travail.

Un objet suivi est une suite continue d'apparences de cet objet à l'intérieur des images du plan vidéo (voir le chapitre précèdent pour plus de détails sur les techniques utilisées pour la localisation et le suivi des objets mobiles). L'objet suivi à un instant t dans le plan vidéo sera désigné par *apparence* intra-plan dans la suite. Nous supposons dans la suite que chaque apparence est représentée par un point unique dans l'espace de descripteurs. Plus formellement, on extrait de l'apparence i de l'objet suivi un descripteur; ce descripteur est décrit sous la forme d'un vecteur v_i de d valeurs réelles; un espace de $\mathbb{R}^d$ est ainsi défini. Une apparence est représentée par un point unique $y_i \in \mathbb{R}^d$. Cette représentation unique d'une apparence dans l'espace de descripteur est valable pour tout descripteur global (voir section 3.2). Les extensions de cette étude pour les descripteurs locaux (les invariants différentiels autour des points d'intérêt, etc.) seront discutées dans les perspectives de ce chapitre.

Si l'objet suivi est mobile et non-rigide – le cas le plus fréquent dans les films vidéo –, plusieurs apparences visuellement différentes peuvent être distinguées pour cet objet. L'apparition progressive intra-plan d'un objet suivi sous une variété de changements d'images (voir l'introduction) produit une **variabilité** dans sa distribution spatiale dans l'espace de descripteurs. Un objet suivi est représenté par un ensemble de descripteurs extraits de ses apparences. Soit $Y = (y_1, ..., y_n)$ la collection de descripteurs (la distribution) de l'objet suivi ayant n apparences, avec y_i le vecteur descripteur de dimension d qui caractérise l'apparence i. La variabilité de l'objet suivi est expliquée par une dispersion des individus y_i dans l'espace $\mathbb{R}^d$. La dispersion est exprimée souvent autour d'un centre. A ce stade on utilise le terme "**degré de variabilité**" pour distinguer entre distribution unimodale et multimodale. La distribution est multimodale si elle est divisible en plusieurs classes différentes. Le degré de variabilité d'une distribution Y dépend du changement significatif de l'apparence intra-plan de l'objet suivi et de la robustesse du descripteur extrait des apparences aux changement d'images.

Aussi, la variabilité d'une distribution d'un objet suivi à travers le temps pourra être bien illustrée par les points de courbures fortes de sa trajectoire spatio-temporelle [5]. Les points de cette trajectoire sont dans un espace de dimension $d+1$, où la dimension temporelle est rajoutée aux d dimensions spatiales. Les points de courbures de la trajectoire signalent des changements significatifs dans l'apparence de l'objet suivi. Notons que le nombre de points de courbures ne correspond pas du tout au degré de variabilité (voir la définition ci-dessus). Ceci est du au fait que beaucoup de points de la trajectoire spatio-temporelle peuvent correspondre à différents segments de la trajectoire sachant qu'ils appartiennent à la même classe d'apparences de l'objet. L'exemple de la figure 3.2.d, détaillé dans l'introduction de ce chapitre, illustre une trajectoire spatio-temporelle bidimensionnelle de l'enfant suivi. Chaque apparence de l'objet à un instant donné est représentée par l'histogramme de couleurs projeté dans le premier plan factoriel. Un rappel de l'ACP appliquée ici pour réduire la dimension de de l'espace de descripteurs est fait dans la section 3.4. Sur cette trajectoire, on voit nettement un point de courbure forte. En conséquence, deux classes d'apparences de l'enfant peuvent être construites : une classe lorsque l'enfant se place entièrement dans la lumière et une deuxième lorsque l'enfant a partiellement disparu dans la partie sombre.

La détermination automatique du degré de variabilité ou plus couramment, le **nombre de classes** d'une distribution d'un objet suivi, est un point fondamental. Ceci est important pour la phase suivante qui porte sur l'appariement des objets suivis basé sur les modèles d'apparences intra-plan des objets. Si on surestime ou sous-estime le nombre de classes d'apparences des objets suivis, le risque de faux appariements entre ces objets augmente. Nous reviendrons sur ce point ultérieurement et en détail dans les deux chapitres suivants.

L'approche que nous proposons pour caractériser un objet suivi consiste à modéliser la distribution de Y comme une fonction de densité de probabilité jointe, $f(y \mid Y, \theta)$, où θ représente l'ensemble de paramètres du modèle f. Nous supposons que f pourra être estimée comme une fonction de densité de mélange gaussien; nous détaillons ce point dans la section suivante. Cette modélisation capture d'une manière efficace la variabilité multimodale de l'objet suivi dans l'espace de descripteurs. Premièrement, les apparences de l'objet suivi sont regroupées dans des classes d'équivalences homogènes. Dans la suite, ces classes sont nommées **classes d'apparences intra-plan**. Deuxièmement, chaque classe d'apparence est caractérisée par quelques paramètres statistiques (la moyenne et la matrice de covariances dans le cas gaussien) estimés à partir de ses éléments (les individus de cette classe). Les modèles f seront utilisés dans l'appariement inter-plan des objets suivis dans une séquence vidéo. Deux objets suivis sont jugés similaires s'ils ont deux classes d'apparences proches au sens d'une distance donnée. Cela justifie notre approche d'identification des classes d'apparences intra-plan d'un objet suivi. Nous détaillons la technique d'usage des modèles d'apparences intra-plan et leurs comparaisons dans les deux chapitres suivants.

3.6 Modélisation par mélange de lois : notre approche

On présente dans cette section la modélisation statistique du problème de la variabilité évoqué précédemment par un mélange de lois. Dans ce travail, on adopte l'hypothèse que la forme de la distribution de chaque classe d'apparences d'un objet suivi suit une loi gaussienne multivariée. Le modèle du mélange est d'abord rappelé dans sa formulation classique. On détaille ensuite le cas des mélanges gaussiens multivariés. Puis on introduit les modèles gaussiens avec contraintes pour éviter la surestimation et/ou l'instabilité de l'estimation des paramètres du mélange, surtout quand le nombre d'apparences d'un objet suivi est relativement petit par rapport à la dimension de l'espace de descripteurs. Enfin, on décrit les méthodes statistiques pour sélectionner automatiquement la structure d'un modèle de mélange, c'est-à-dire le modèle gaussien et le nombre de composantes (classes) de la densité du mélange. Le nombre de classes ou le degré de variabilité diffère d'un objet suivi à l'autre (voir la section précédente).

3.6.1 Modèle de mélange

En classification automatique, on cherche généralement à détecter un regroupement des observations en J classes homogènes et distinctes les unes des autres. Aussi, lorsqu'on cherche à définir un modèle probabiliste, il paraît naturel de supposer que les observations ont été générées à partir de J distributions homogènes, c'est-à-dire concentrées autour de leur valeur centrale et se chevauchant peu. Les modèles de mélange constituent à cet égard un cadre simple et adapté pour de tels problèmes de classification.

3.6.1.1 Caractérisation d'une distribution mélange

Dans un modèle de mélange les observations $\mathbf{y} = (y_1, ..., y_n)$ sont supposées être des réalisations indépendamment et identiquement distribués (i.i.d.)de n vecteurs aléatoires $\mathbf{Y} = (Y_1, ..., Y_n)$. La distribution du vecteur aléatoire parent est supposée être un mélange de J distributions, caractérisées par des fonctions de densité de probabilité $\varphi(. \mid \alpha_j)$, $1 \leq j \leq J$. Ces J distributions ont pour paramètres $\alpha_1, ..., \alpha_J$, et sont mélangées selon des proportions respectives $p_1, ..., p_J$, avec $0 < p_j < 1$ et $\sum_{j=1}^{J} p_j = 1$. La densité de probabilité en un point $y \in I\!R^d$ est ainsi donnée par :

$$f(y|\theta) = \sum_{j=1}^{J} p_j \varphi(y|\alpha_j) \tag{3.6}$$

Le vecteur θ dénote les paramètres inconnus du mélange. Dans le cas le plus général, $\theta = (p_1, \ldots, p_J, \alpha_1, \ldots, \alpha_J)$.

3.6.1.2 Mélanges gaussiens

Lorsque les données sont quantitatives, c'est-à-dire à valeurs réelles continues, et en absence de connaissances particulières concernant les distributions du mélange, il est courant de supposer que chaque classe suit une distribution gaussienne multivariée. Dans cette situation multivariée, la densité de la composante j en un point $y \in I\!R^d$ est donnée par

$$\varphi(y \mid \mu_j, \mathbf{\Sigma}_j) = \frac{1}{(2\pi)^{d/2} \mid \mathbf{\Sigma}_j \mid^{1/2}} exp\left(-\frac{1}{2}(y-\mu_j)'\mathbf{\Sigma}_j^{-1}(y-\mu_j)\right) \quad (3.7)$$

où $\mid (.) \mid$ dénote le déterminant de la matrice (.). La distribution de chaque classe est paramétrée par son vecteur moyenne μ_j $(d \times 1)$ et sa matrice de variance Σ_j ($d \times d$ symétrique définie positive).

Les paramètres à estimer pour le mélange $f(x \mid \theta)$ sont donc :

$$\theta = (\theta_j) \text{pour } j = 1, \ldots, J \text{ et avec } \theta_j = (p_j, \mu_j, \Sigma_j).$$

La figure 3.3 montre le type d'allure que peuvent prendre des données distribuées selon un modèle de mélange gaussien, respectivement dans $I\!R$ et $I\!R^2$. Le figure 3.3.a illustre les données x dans $I\!R^2$ $(n = 114)$. La figure 3.3.b montre la partition obtenue par l'algorithme EM (voir section 3.6.5) où le nombre de classes (fixé à 5) et le modèle gaussien général (voir section 3.6.2) ont été fixés a priori. La densité du mélange correspondante, dans $I\!R$ et $I\!R^2$ est illustrée dans la figure 3.3.c et 3.3.d respectivement.

3.6.2 Modèles gaussiens avec contraintes

3.6.2.1 Motivation pour ce travail

La réduction de la dimension de l'espace de descripteurs par une ACP est une étape fondamentale pour mieux conditionner l'estimation des paramètres du mélange gaussien. Mais la dimension initiale des descripteurs globaux est généralement élevée (64, 128, etc.), et la dimension retenue du nouvel espace de descripteurs après l'ACP, pour une qualité de représentation de 90% par exemple, reste encore grande vis-à-vis du petit nombre d'apparences intra-plan de quelques objets suivis dans la vidéo (cas où la durée du plan vidéo est d'une seconde). On reste alors face à un problème de surparamétrisation. Une solution pour ce problème consiste à réduire la complexité de l'approche de modélisation de la variabilité intra-plan. Ceci pourra être fait par réduction du nombre de paramètres à estimer surtout les coefficients de la matrice de variance Σ. Nous introduisons dans la suite la notion des modèles gaussiens *parcimonieux*. Lorsque l'on modélise la distribution d'un échantillon par un mélange gaussien différentes hypothèses peuvent être adoptées en fonction du problème traité. Si l'on utilise le modèle le plus général, on laisse tous les paramètres varier : le vecteur de paramètres inconnu est θ. Cependant, il est parfois avantageux en pratique de se baser sur un modèle plus contraint, en supposant par exemple que les matrices de variance sont identiques. La section 3.6.2.3 décrit les 7 modèles gaussiens dont nous nous servis.

Avoir moins de paramètres à estimer rend leur estimation plus fiable lorsque l'on dispose de peu d'observations. En principe, l'utilisateur fixe ces contraintes d'après les connaissances dont il dispose sur le problème modélisé. Mais, dans ce travail il s'agit d'appliquer l'approche de modélisation intra-plan sur la distribution de chaque objet suivi séparément. Cette distribution a probablement une structure différente d'un objet suivi à

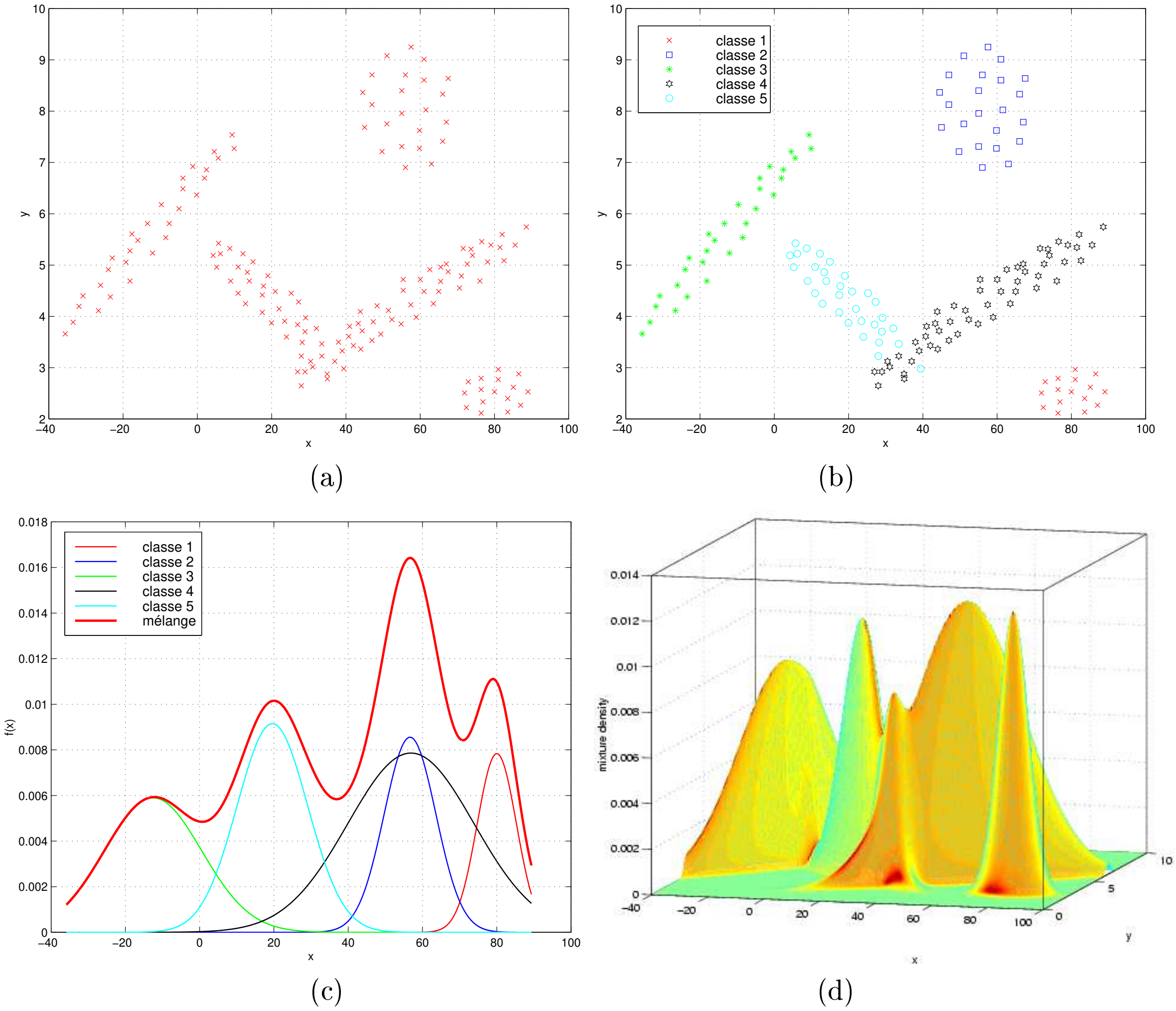


FIG. 3.3: *(a) illustration de données dans $\mathbb{R}^2$; (b) la partition correspondante en cinq classes; (c) la densité du mélange dans $\mathbb{R}$ et (d) dans $\mathbb{R}^2$.*

un autre. Donc, fixer un modèle gaussien avec contraintes pour toutes les distributions des objets suivis, ou bien laisser "l'auteur de la vidéo hyperliée" de la vidéo choisir un modèle manuellement pour chaque distribution, sont deux solutions à éviter. En effet, il s'avère très utile dans ce cas d'utiliser une stratégie qui permette de sélectionner automatiquement le modèle gaussien modélisant le mieux la distribution d'un objet suivi. Pour ce faire, la section 3.6.6 introduit quelques critères implementés dans ce travail.

3.6.2.2 Les 4 modèles gaussiens de bases

Les connaissances a priori les plus classiques dont l'utilisateur dispose portent souvent sur les proportions p_j et les matrices de variances des classes Σ_j. Ainsi, les proportions peuvent être supposées identiques, et n'ont pas besoin d'être estimées ($p_1 = \ldots = p_J =$

$1/J$). Concernant les matrices de variance, quatre modèles sont couramment utilisés (cf. [34] [15]) :

1. Modèle linéaire sphérique : $\Sigma_1 = \ldots = \Sigma_J = \sigma^2\mathbf{I}$, où $\mathbf{I}$ dénote la matrice identité $d \times d$). Dans chaque classe les variables sont supposées indépendantes et dans toutes les classes les variables ont la même variance inconnue $\sigma^2 > 0$.

2. Modèle linéaire diagonal : $\Sigma_1 = \ldots = \Sigma_J = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ avec $(\sigma_1^2, \ldots, \sigma_d^2)$ inconnus représentent les éléments diagonaux de la matrice diagonale $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Dans chaque classe les variables sont supposées indépendantes.

3. Modèle linéaire général : $\Sigma_1 = \ldots = \Sigma_J = \Sigma$. A l'intérieur de chaque classe, les variables peuvent être corrélées, et les différentes classes partagent la même matrice de variance inconnue Σ (matrice $d \times d$ symétrique définie positive).

4. Modèle quadratique : aucune contrainte sur les Σ_j. Les variables peuvent être corrélées à l'intérieur de chaque classe et les classes peuvent posséder des structure de covariance différente.

Les trois premiers modèles sont dits linéaires car les séparations entre les classes sont des hyper-plans dans $\mathbb{R}^d$. Ceci est dû au fait que, dans ces trois modèles, les classes ont la même matrice de variance.

3.6.2.3 Les modèles parcimonieux

Banfield et Raftery [6] proposent une décomposition spectrale de la matrice de variance afin de raffiner les quatre modèles gaussiens présentés ci-dessus. Cette décomposition permet de spécifier la structure de variance des classes par des paramètres intuitifs, interprétables d'un point de vue aussi bien statistique que géométrique. Nous présentons ici la méthode adoptée dans ce travail, la décomposition de Celeux et Govaert [22]. Cette décomposition modifie un peu celle de Banfield et Raftery en facilitant le calcul des estimateurs du maximum de vraisemblance. L'idée de base est d'écrire chaque matrice de variance sous la forme

$$\Sigma_j = \lambda_j D_j A_j D_j'$$

où

- le paramètre scalaire positif $\lambda_j = \det(\Sigma_j)^{1/d}$, appelé *volume* de la classe j, indique la dispersion de la classe dans $\mathbb{R}^d$; ce paramètre, homogène à la variance des variables de la classe est égal à la moyenne géométrique des valeurs propres de Σ_j;

- la matrice $d \times d$ orthogonale D_j est constituée en colonne des vecteurs propres de Σ_j. Appelée *orientation*, elle représente géométriquement les axes de l'ellipsoïde de dispersion de la classe j, et statistiquement le degré de corrélation entre les variables dans la classe j;

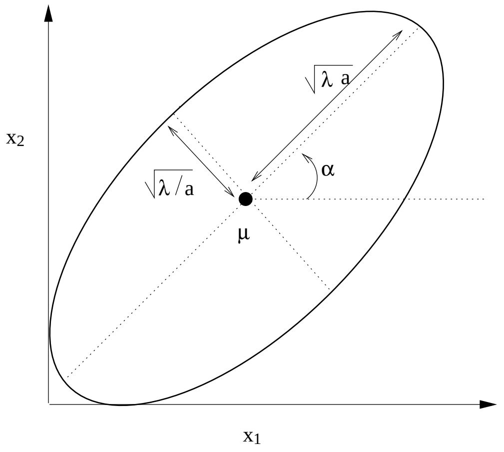


FIG. 3.4: *Décomposition spectrale de la matrice de variance : relation entre l'ellipse de dispersion et les paramètres de volume, de forme et d'orientation.*

– La matrice diagonale $A_j = \text{diag}(a_{j1}, \ldots, a_{jd})$ est formée par les valeurs propres de Σ_j, rangées par ordre décroissant et normalisées de sorte à ce que $\det(A_j) = \prod_{\ell=1}^{d} a_{jd} = 1$. Appelé paramètre de *forme*, A_j indique l'allongement relatif de l'ellipsoïde de dispersion de la classe j le long de ses axes. D'un point de vue statistique, les a_{jd} peuvent être interprétés comme les dispersions relatives des variables après le changement de base induit par la matrice d'orientation D_j.

Selon si ces trois paramètres entre les classes varient ou sont égaux, selon les formes supposées sphériques ($A_j = \mathbf{I}$) ou non et selon si les orientations sont parallèles aux axes (D_j = matrice de permutation de la base canonique) ou obliques, on obtient 14 modèles gaussiens, du plus parcimonieux au moins parcimonieux [22]. Ces 14 modèles peuvent être regroupés en 3 familles :

1. famille sphérique : $A_j = \mathbf{I}$, ce qui donne une matrice $\Sigma_j = \lambda_j \mathbf{I}$ sphérique; l'orientation D_j n'a donc aucune influence dans ce cas.

2. famille diagonale : la matrice de forme A_j est quelconque, et la matrice d'orientation D_j définit une permutation de la base canonique, ce qui donne une matrice diagonale; on note que $B_j = D_j A j D'_j$ la matrice normalisé (diagonale) avec $det(B) = 1$;

3. famille générale : forme et orientation quelconques, la matrice Σ_j n'est pas forcément diagonale; on note $C_j = D_j A_j D'_j$ la matrice normalisée.

La liste complète de ces 14 modèles gaussiens ainsi que l'étape de maximisation M de l'algorithme EM sont détaillées dans l'article de [22]. Dans ce travail, nous nous sommes restreints à la moitié de ces 14 modèles résumés dans le tableau 3.1. Ce choix est motivé principalement par deux choses : (1) ces modèles, partagés entre les trois grandes familles de modèles gaussiens soulignées ci-dessus, ont des complexités variables (faible, moyenne et forte), ils peuvent être donc adaptés à une distribution d'un objet suivi avec un peu ou beaucoup d'apparences intra-plan; (2) le temps de la sélection automatique de la structure d'un mélange gaussien (section 3.6.6) sera réduit au moins au moitié et donc une modélisation de la variabilité est faisable dans une durée raisonnable.

Famille	Modèles	Nombre de paramètres Cas général	J = 2 et d = 10
Sphérique	$[\lambda I]$	$\alpha + 1$	22
	$[\lambda_j I]$	$\alpha + J$	23
Diagonale	$[\lambda_j B]$	$\alpha + d + J - 1$	32
	$[\lambda_j B_j]$	$\alpha + Jd$	41
Générale	$[\lambda_j C]$	$\alpha + \beta + J - 1$	76
	$[\lambda C_j]$	$\alpha + J\beta - (J - 1)$	130
	$[\lambda_j C j]$	$\alpha + J\beta$	131

TAB. 3.1: *Liste des 7 modèles gaussiens retenus;* $\beta = (d(d + 1)/2)$ *et* $\alpha = Jd + J - 1$.

Dans le tableau 3.1 chaque modèle est désigné par une notation abrégée indiquant quels paramètres sont identiques ou différents entre les classes. Ainsi, dans cette notation, le modèle imposant un même volume entre toutes les classes est noté $[\lambda]$, le modèle permettant des volumes différents entre classes est noté $[\lambda_j]$. Aussi, $[\lambda_j D_j A D_j']$ désigne le modèle à volume et orientation libres entre les classes mais à forme identique. les quatres modèles de variance classiques cités plus haut seraient respectivement désignés par $[\lambda I]$ (modèle linéaire sphérique), $[\lambda I]$ (linéaire diagonal), $[\lambda C]$ (linéaire général), $[\lambda_j C_j]$ (quadratique). Dans [22], on souligne par exemple que le modèle $[\lambda_j I]$, matrices de variances sphériques de volume différents $\Sigma_j = \sigma_j I$, requiert l'estimation de seulement J paramètres pour les variances, contre $J \times d \times (d + 1)/2$ paramètres avec le modèle $[\lambda_j C_j]$.

3.6.2.4 Ellipse de dispersion

Pour une distribution gaussienne Y dans $I\!R^2$ de centre μ et de matrice de variance $\Sigma = \lambda DAD'$, on peut noter que le lieu des points $y \in I\!R^2$ ayant une densité est caractérisé par une équation de la forme

$$(y - \mu)'\Sigma^{-1}(y - \mu) = \Delta \tag{3.8}$$

où Δ est une constante positive qui dépend du niveau de densité choisi. L'équation 3.8 définit en fait l'ensemble des points y situés à une distance de Mahalanobis de μ égale à Δ, distance pondérée par Σ. On verra plus loin que la courbe ainsi définie contient comme donnée une proportion p_Δ des points issus de la distribution Y.

L'équation 3.8 définit une ellipse centrée en μ. En effet, cette équation s'écrit aussi (D étant orthogonale, $D^{-1} = D'$):

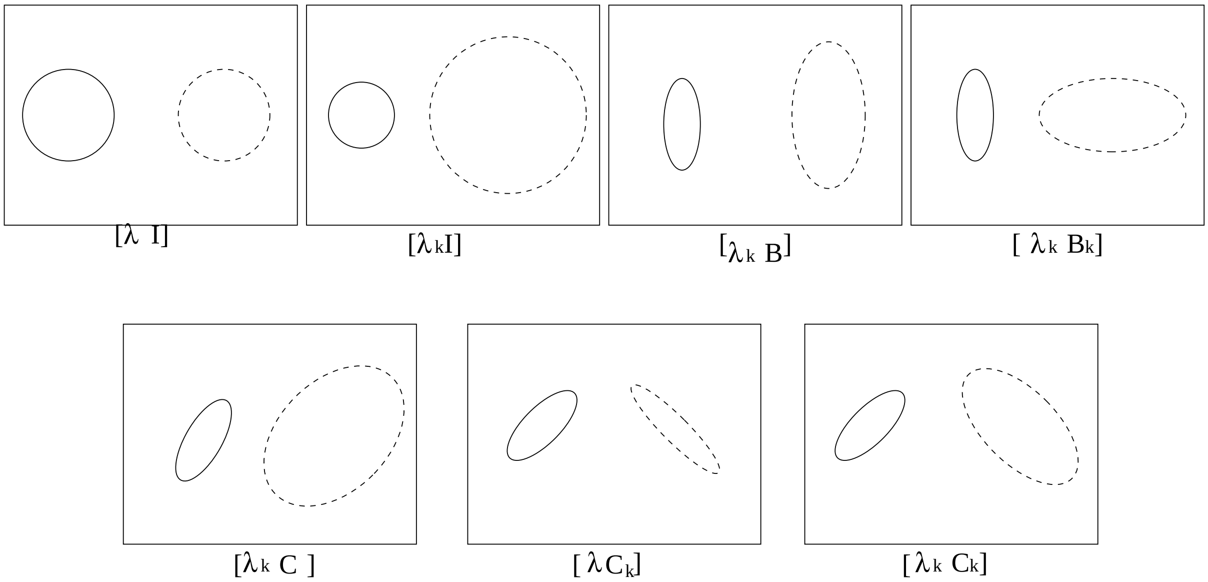


FIG. 3.5: *Illustration des ellipses de dispersion des 7 modèles gaussiens utilisés pour modéliser des données de deux composantes gaussiennes. Chaque couple d'ellipses est associé à un seul modèle gaussien.*

$$[D'D(y-\mu)]'D'(\lambda A)^{-1}D(y-\mu) = \Delta$$

$$[D(y-\mu)]'(\lambda A)^{-1}D(y-\mu) = \Delta.$$

On peut faire le changement de variable

$$\hat{y} \approx D(y-\mu)$$

ce qui revient se placer dans le repère centré en μ et orienté suivant les axes définis par D. Par ailleurs, dans $I\!R^2$, la matrice de forme A a pour expression

$$A = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$$

avec $a \in I\!R$. L'équation ci-dessus peut donc s'écrire

$$\hat{y}' \begin{pmatrix} \frac{1}{\lambda a} & 0 \\ 0 & \frac{a}{\lambda} \end{pmatrix} \hat{y} = \Delta$$

$$\frac{(\hat{y_1})^2}{\lambda a} + \frac{(\hat{y_2})^2}{\frac{\lambda}{a}} = \Delta$$

ce qui définit une ellipse centrée en μ et orientée suivant les axes du nouveau repère. Les grand et petit axes de cette ellipse ont pour longueurs respectives $\sqrt{\Delta \lambda a}$ et $\sqrt{\Delta \frac{\lambda}{a}}$.

3.6.3 Estimation du paramètrage du mélange

L'estimation consiste à donner des valeurs approchées aux paramètres d'un échantillon à l'aide de n observations issues de cet échantillon. Différentes méthodes d'estimation du paramètre du mélange gaussiens multivariés θ peuvent être appliquées: maximum de vraisemblance (MV), estimation bayésienne, théorie du codage, etc. Le paragraphe 3.6.3.2 décrit brièvement le principe d'estimation par MV et l'estimation bayésienne. Nous détaillons ci-après l'estimation par MV, qui sera utilisée dans notre approche d'identification des classes d'apparences intra-plan. Plus précisément, on se focalise sur ce cas automatique où on dispose seulement des observations $y_1, ..., y_n$, les vecteurs descripteurs des apparences d'un objet suivi.

Définition 3.6.1 *Estimateur-Estimation : On appelle estimateur T du paramètre θ toute fonction ϕ de l'échantillon $Y_1, ..., Y_n$, $T = \phi(Y_1, ..., Y_n)$. La réalisation de cet estimateur, $t = \phi(y_1, ..., y_n)$, est appelée estimation et sera notée dans la suite par $\hat{\theta}$.*

3.6.3.1 Conditions d'identifiabilité du mélange

Pour estimer les paramètres d'un modèle -le mélange dans notre cas- celui-ci doit-être *identifiable.* L'indentifiabilité signifie que le modèle, f, possède une décomposition unique en terme de composantes, φ_j. Autrement dit, dans un modèle identifiable, si deux mélanges produisent une même fonction de densité, on a :

$$\sum_{k=1}^{K} p_k \varphi(.|\alpha_k) = \sum_{j=1}^{J} p'_j \varphi(.|\alpha'_j)$$

d'où, compte-tenu de l'unicité de la décomposition :

$$\begin{cases} K = J \\ \text{et} \\ \forall k, p_k = p'_k \text{ et } \varphi(.|\alpha_k) = \varphi(.|\alpha'_k). \end{cases}$$

Yakowitz et Spragins présentent dans [126] une étude approfondie sur la caractérisation des mélanges identifiables. Ils montrent que les mélanges gaussiens sont en particulier identifiables. Au contraire, les mélanges de lois uniformes consituent un cas typique de modèle non identifiable. Pour le constater, prenons l'exemple de la densité uniforme $U_{[0;1]}$ sur l'intervalle $[0, 1]$. Cette densité peut être décomposée en deux densités uniformes du type $U_{[0,p]}$ et $U_{[p,1]}$ en proportions respectives p et $1 - p$, $\forall p \in]0, 1[$.

3.6.3.2 Principes d'estimation

Estimation par maximum de vraisemblance Le principe du maximum de vraisemblance est l'un des plus utilisés pour estimer le paramètre θ d'une distribution f en se basant sur une réalisation y d'un échantillon Y de cette distribution. Cette démarche consiste à chercher l'expression de paramètre θ qui maximise la probabilité $f(y \mid \theta)$ d'observer x avec comme vecteur de paramètres θ. Pour une réalisation y fixée, la probabilité

$f(y \mid \theta)$ est une fonction de paramètre appelée vraisemblance. En pratique, on cherche plutôt à maximiser le logarithme de la vraisemblance, noté $L(\theta)$, ce qui est équivalent et conduit généralement à des calculs plus simples. On cherche donc :

$$\hat{\theta}_{MV}(y) = \arg\max_{\theta} \Big(L(\theta) = \log f(y \mid \theta) \Big)$$

La log-vraisemblance du paramètre θ d'un mélange gaussiens $f(y \mid \theta)$ (équations 3.6, 3.7) est donnée par :

$$L(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{J} p_j \varphi(y_i | \mu_j, \Sigma_j) \right\}. \tag{3.9}$$

D'autres formulations du MV existent lorsque les données et leurs labels sont partiellement ou complétement connus.

Estimation bayésienne. Dans l'approche bayésienne de l'inférence statistique, l'estimation de paramètre θ d'un modèle est vue comme un problème de *décision* concernant le vrai paramètre inconnu θ au vu de données $y = \{y_1, \dots, y_n\}$. On se donne pour cela :

- l'estimateur $l(\hat{\theta} \mid \theta)$ qui est une fonction de coût indiquant le coût d'une décision $\hat{\theta}$ si la vrai valeur est θ; $l(\hat{\theta} \mid \theta)$ associe a chaque décision $\hat{\theta} \in$ {espace des décisions} $\times$ {espace des paramètres} une valeur réelle positive;
- et une loi a priori des paramètres $\pi(\theta)$.

L'estimateur de Bayes $\hat{\theta}$ de θ est celui qui minimise le risque a posteriori, donné par

$$T(\hat{\theta} \mid \theta) = E[l(\hat{\theta} \mid \theta) \mid y] = \int l(\hat{\theta} \mid \theta) \pi(\theta \mid y) d\theta$$

avec $\pi(\theta \mid y)$ dénote la loi a posteriori des paramètres qui est déterminée par la connaissance de la loi a priori et de la vraisemblance (formule de Bayes) :

$$\pi(\theta \mid y) = \frac{\pi(\theta)\pi(y \mid \theta)}{\pi(y)} = \frac{\pi(\theta)\pi(y \mid \theta)}{\int \pi(\theta)\pi(y \mid \theta) d\theta}$$

La fonction de coût la plus utilisée est le coût quadratique, $l(\hat{\theta} \mid \theta) = (\hat{\theta} - \theta)'(\hat{\theta} - \theta)$, qui donne généralement lieu à des calculs plus simples. La minimisation du risque a posteriori $T(\hat{\theta} \mid \theta)$ sur $\hat{\theta}$ est résolue explicitement dans ce cas quadratique, mais nécessite la connaissance de la loi a posteriori. Malheureusement, en mélange, la loi a posteriori, quoique explicite, n'est pas calculable pour une taille d'échantillon raisonnable. Ceci est confirmé par le contre exemple traité par Robert [103] (exemple 9.5 du chapitre 9). Récemment, des méthodes de simulations des réalisations, connues sous le nom MCMC (*Markov Chain Monte Carlo*) sont proposées pour palier ce problème de calcul direct de la loi a posteriori. Nous nous servirons des méthodes de Monte Carlo dans l'évaluation de la distance de Kullback calculée entre deux distributions gaussiennes (voir section 5.3).

3.6.4 Maximum de vraisemblance par l'algorithme EM

3.6.4.1 Information manquante pour le mélange

En classification automatique, l'idée de base est de trouver un ensemble des variables manquantes, $z = (z_1, \ldots, z_n)$, qui relient les observations $y = (y_1, \ldots, y_n)$ à J classes inconnues. Les variables manquantes z représentent les labels ou les classes des observations : y_i est classé dans la j^{eme} classe si et seulement si $z_i = j$. Soit $x = (y, z)$ les données complètes. Le principe du maximum de vraisemblance pour ces données complètes x donnerait lieu à des calculs simples en général. En effet, la vraisemblance du θ pour x est alors

$$\begin{aligned} L(x,\theta) = \log f(y,z \mid \theta) &= \sum_{i=1}^{n} \log \Big(p_{z_i} \varphi(y_i, \alpha_{z_i}) \Big) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log \Big(p_j \varphi(y_i, \alpha_j) \Big) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log p_j + \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log \varphi(y_i, \alpha_j)}_{\log f(y_i, \mid \theta)} \end{aligned} \tag{3.10}$$

avec $z_{ij} = 1$ ou 0 selon que $z_i = j$ ou non.

La recherche des paramètres des distributions pourrait se faire en maximisant séparement la vraisemblance $f(y_i, \mid \theta)$ à l'intérieur de chaque classe. Cette maximisation ne pose généralement pas de difficulté dans la mesure où les composants d'un mélange ont la plupart du temps des distributions simples.

L'algorithme *Expectation-Maximization* (EM) est précisement conçu pour ce type de problème d'estimation, où le principe du maximum de vraisemblance pour les données complètes x donnerait des calculs simples, mais où l'on n'observe en fait qu'une partie y de ces données [31].

3.6.4.2 Algorithme itératif

Il s'agit d'une procédure d'estimation itérative, qui commence avec une valeur initiale des paramètres θ^0. Chaque itération consiste ensuite à calculer les nouveaux paramètres θ^{m+1} à partir de ceux de l'itération précédente θ^m de façcon à maximiser une fonction notée $Q(\theta, \theta^m)$, ainsi définie :

$$Q(\theta, \theta^m) = \Big\langle \log f(y, z \mid \theta) \mid y, \theta^m \Big\rangle. \tag{3.11}$$

L'expression 3.11 s'interprète comme l'espérance de la log-vraisemblance complète, espérance prise sur la distribution a posteriori des données manquantes z connaissant les observations x et basée sur les anciens paramètres θ^m.

Chaque itération peut se décomposer en deux étapes [31] :

- Etape **E** : calculer les composantes de l'espérance $Q(\theta, \theta^m)$ qui ne dépendent pas de θ.

- Etape **M** : mettre à jour les paramètres par : $\theta^{m+1} = \arg\max_\theta Q(\theta, \theta^m)$

La croissance de la vraisemblance à chaque itération de cet algorithme connu sous le nom EM (*Expectation-Maximization*) est demontrée dans l'annexe A.

3.6.5 EM et ses variantes appliqués au mélange gaussien

L'algorithme EM est particulièrement adapté pour estimer les paramètres d'un mélange fini de distributions [98] [21]. Dans cette situation, les informations manquantes z sont les classes des observations et les paramètres du modèle sont θ. A l'itération $m+1$, l'espérance à maximiser s'écrit dans ce cas

$$Q(\theta, \theta^m) = \sum_{i=1}^{n} \sum_{j=1}^{J} \underbrace{\left\langle z_{ij} \mid x, \theta^m \right\rangle}_{t_{ij}^m} \log(p_j f(x_i \mid \alpha_j))$$

A l'étape E, on calcul les parties de $Q(\theta, \theta^m)$ qui ne dépendent pas de θ, soit ici, pour $1 \leq i \leq n$ et $1 \leq j \leq J$

$$t_{ij}(\theta^m) = \frac{p_j^m \varphi(x_i \mid \mu_j^m, \Sigma_j^m)}{\displaystyle\sum_{\ell=1}^{J} p_\ell^m \varphi(x_i \mid \mu_\ell^m, \Sigma_\ell^m)}. \tag{3.12}$$

c'est à dire les probabilités d'appartenance a posteriori des observations aux classes connaissant les données x et en se basant sur les paramètres θ^m de l'itération précédente.

L'étape M consiste à chercher

$$\theta^{m+1} = \arg\max_\theta \sum_{\ell=1}^{J} t_{ij}(\theta^m) \log p_j + \sum_{\ell=1}^{J} t_{ij}(\theta^m) \log f(x_i \mid \varphi_j)$$

c'est-à-dire

$$p_j^{m+1} = \arg\max_{p_j} \sum_{j=1}^{J} n_j^{m+1} \log p_j \tag{3.13}$$

$$\forall \; 1 \leq j \leq J, \;\; \varphi_j^{m+1} = \arg\max_{\varphi_j} \sum_{i=1}^{n} t_{ij}(\theta^m) \log f(x_i \mid \varphi_j) \tag{3.14}$$

où

$$n_j^{m+1} = \sum_{i=1}^{n} t_{ij}^m(\theta^m)$$

peut s'interpréter comme l'espérance du nombre d'observations appartenant à la classe j. La résolution de l'équation 3.13 sous la contrainte $\sum_{j=1}^{J} p_j = 1$ donne, pour $1 \leq j \leq J$,,

$$p_j^{m+1} = \frac{n_j^{m+1}}{n}.$$

Quant à la recherche des paramètres des classes, chacun de ces J sous problèmes peut être vu comme une maximisation de la vraisemblance des paramètres φ_j à l'intérieur de la classe j, en pondérant chaque individu x_i par son degré d'appartenance t_{ij}^m à cette classe. Par exemple, dans le cas du mélange gaussiens le plus général, les formules obtenues pour recalculer les paramètres peuvent être interpétées comme des estimateurs du maximum de vraisemblance de chaque distribution en pondérant les individus par t_{ij}^m :

$$\mu_j^{m+1} = \frac{1}{n_j^{m+1}} \sum_{i=1}^{n} t_{ij}^m(\theta^m) y_i \tag{3.15}$$

$$\Sigma_j^{m+1} = \frac{1}{n_j^{m+1}} \sum_{i=1}^{n} t_{ij}^{m+1}(\theta^m)(y_i - \mu_j^{m+1})(y_i - \mu_j^{m+1})^\mathsf{T} \tag{3.16}$$

La structure d'un lancer de EM est récapitulée dans la figure 3.6.5 pour le cas du modèle de mélange gaussien le plus général.

Convergence. La convergence de EM (voir annexe A) a été établie dans le cadre des modèles de mélange par Redner et Walker [98] sous des conditions asymptotiques assez peu contraignantes. Leur théorème met en relief la nécessité de connaître le nombre de classes et l'importance d'une initialisation qui ne soit pas trop éloignée des vrais paramètres. Redner et Walker soulignent de plus que la convergence de EM, de nature linéaire, est d'autant plus rapide que les composants sont bien séparés. Souvent la solution produite par EM dépend fortement de la position initiale, et l'algorithme EM peut converger vers un palier (non maximum) de la vraisemblance ou rester très longtemps sur un tel palier. De plus, même lorsque EM converge vers un maximum de la vraisemblance, celui-ci n'est que local. Or, dans le cadre des modèles de mélange, la surface de la vraisemblance présente souvent plusieurs maximum locaux [83].

Initialisation. Différentes techniques peuvent être envisagées pour atténuer la dépendance de l'algorithme EM à la position initiale. La tactique la plus fréquement utilisée est similaire à celle décrite pour la méthode de centres-mobiles : on lance plusieurs fois l'algorithme EM de différentes positions initiales θ^0 choisies au hasard, et l'on retient la solution $\hat{\theta}$ qui donne la plus grande vraisemblance $L(\theta)$. Cette tactique est simple à mettre en place, mais elle peut être coûteuse en temps de calcul dans les cas où l'algorithme converge lentement.

Par ailleurs, d'autres techniques itératives d'optimisation peuvent être employées pour tenter de palier la lenteur de l'algorithme. Cependant, nous n'avons pas étudié ce point durant cette thèse. A titre d'exemple, on peut citer la méthode de Newton, qui consiste à mettre à jour le paramètre θ en appliquant la formule suivante :

$m \leftarrow 0$
Initialiser les proportions, centres et variances a une valeurs arbitraire θ^0
répéter
{ Etape **E**: calculer les probabilités d'appartenance t_{ij}^m}
pour i de 1 à n, j de 1 à J faire
$$t_{ij}(\theta^m) \leftarrow \frac{p_j^m \varphi(x_i|\mu_j^m, \Sigma_j^m)}{\sum_{\ell=1}^{J} p_\ell^m \varphi(x_i|\mu_\ell^m, \Sigma_\ell^m)}$$
{ Etape **M**: recalculer les paramètres θ^{m+1}}
pour j de 1 à J faire
$$n_j^{m+1} \leftarrow \sum_{i=1}^{n} t_{ij}^m(\theta^m); \quad p_j^{m+1} \leftarrow \frac{n_j^{m+1}}{n}$$
$$\mu_j^{m+1} \leftarrow \frac{1}{n_j^{m+1}} \sum_{i=1}^{n} t_{ij}^m(\theta^m) y_i$$
$$\Sigma_j^{m+1} \leftarrow \frac{1}{n_j^{m+1}} \sum_{i=1}^{n} t_{ij}^{m+1}(\theta^m)(y_i - \mu_j^{m+1})(y_i - \mu_j^{m+1})^\mathsf{T}$$
$m \leftarrow m+1$
jusqu'à $t^m = t^{m-1}$ *ou* [ou $m = mMAX$]
$\hat{\theta} \leftarrow \theta^m$

FIG. 3.6: *Un lancer de l'algorithme EM*

$$\theta^{m+1} = \theta^m - H_L(\theta^m)^{-1} \nabla_\theta L(\theta^m)$$

où $H_L(\theta^m)^{-1}$ désigne la matrice hessienne de $L(\theta)$ en θ^m. Cette procédure a une vitesse de convergence quadratique, mais elle requiert des calculs importants pour inverser la matrice Hessienne.

Plusieurs tests de convergence de l'algorithme EM peuvent être envisagés. Une stratégie simple consiste à tester la stabilisation des probabilités d'appartenances. Une façon de procéder est de vérifier si la plus grande différence entre les probabilités d'appartenance est inférieure à un certain seuil :

$$\max_{i,j} \mid t_{ij}^{m+1} - t_{ij}^m \mid \; < \epsilon$$

avec par exemple $\epsilon = \frac{0.01}{J}$. Outre sa simplicité de mise en oeuvre, ce test donne lieu à une interprétation intuitive du seuil ϵ. Une stratégie alternative est de tester la stabilisation du critère de vraisemblance, par une condition du type

$$\left| \frac{L(\theta^{m+1}) - L(\theta^m)}{L(\theta^m)} \right| \; < \epsilon$$

avec un seuil ϵ arbitrairement choisi, par exemple $\epsilon = 10^{-6}$. De plus, une condition d'arrêt supplémentaire $m = mMax$ est souvent utilisée afin de ne pas laisser l'algorithme EM séjourner pendant de nombreuses itérations dans un col de la vraisemblance. Ceci peut être utile si l'on envisage de redémarrer l'algorithme un certain nombre de fois et de garder le meilleur résultat. Notons que dans notre implémentation de l'algorithme EM, nous couplons ces deux derniers critères pour tester la convergence. Un dernier test de convergence consiste à tester la stabilisation des paramètres en évaluant la distance entre ces paramètres de deux itérations successives. Cependant ce calcul est relativement complexe à mettre en oeuvre.

Variante CEM. Dans ce travail nous utilisons une variante de EM, connue sous le nom de CEM (classifiante EM) qui converge beaucoup plus rapidement (d'un facteur de 10 au moins) que EM ([21]). L'algorithme CEM cherche à maximiser le critère de vraisemblance classifiante donné par :

$$L_c(z,\theta) = \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} \log\left\{p_j \varphi(y_i|\mu_j,\Sigma_j)\right\}. \tag{3.17}$$

Il possède une étape intermédiaire C entre les deux étapes de l'algorithme classique EM (algorithme 3.6.5). Cette étape consite à calculer, à partir des probabilités d'appartenance t_{ij}, une classification z par la règle de maximum a posteriori. La figure 3.6.5 résume un lancer de cet algorithme dans le cas d'un modèle gaussien le plus général.

3.6.6 Structures en compétition

L'algorithme EM et ses variantes appliquées sur les mélanges exigent en général que l'utilisateur fixe au préalable le nombre de classes J et les contraintes à imposer sur les paramètres. Or, face à un problème réel et plus particulièrement à notre usage des mélanges gaussiens pour modéliser l'apparence intra-plan d'un objet suivi, on n'a aucune information a priori sur le degré de variabilité de cet objet c'est-à-dire le nombre de classes d'apparences intra-plan de l'objet. En plus, le degré de variabilité d'un objet peut différer d'un objet à un autre. De ce fait, et aussi pour les raisons discutées dans la section 3.6.2.1 il parait très utile d'utiliser des techniques qui permettent de déterminer automatiquement la structure du mélange gaussien c'est-à-dire le meilleur nombre de composantes gaussiennes et le modèle gaussien le plus approprié aux données modélisées.

De nombreuses approches ont été développées pour sélectionner automatiquement la structure de mélange la plus adéquate pour un échantillon donnée : tests d'hypothèse, facteur de Bayes, critères d'information et le critère de vraisemblance classifiante [71] [28] [23] [10]. Le lecteur intéressé pourra se référer à la thèse de Biernacki [9], qui consacre à ce sujet une étude expérimentale approfondie sur des données simulées. Nous décrivons plus particulièrement ici les critères d'information et le critère de vraisemblance classifiante dont nous nous sommes servis dans ce travail pour leur fondement théorique justifiable et leur caractère de simplicité de mise en oeuvre.

$m \leftarrow 0$
Initialiser les proportions, centres et variances a une valeurs arbitraire θ^0
répéter
{ Etape **E**: calculer les probabilités d'appartenance t_{ij}^m}
pour i de 1 à n, j de 1 à J faire
$$t_{ij}(\theta^m) \leftarrow \frac{p_j^m \varphi(x_i|\mu_j^m, \Sigma_j^m)}{\sum_{\ell=1}^{J} p_\ell^m \varphi(x_i|\mu_\ell^m, \Sigma_\ell^m)}$$
{ Etape **C**: calculer la classification z^mpar MAP }
pour i de 1 à n, j de 1 à J faire
$z_i^m \leftarrow e_j$avec
$j = \arg\max_j t_{ij}^m$
{ Etape **M**: recalculer les paramètres θ^{m+1}}
pour j de 1 à J faire
$$n_j^{m+1} \leftarrow \sum_{i=1}^{n} z_{ij}^m(\theta^m);$$
$$p_j^{m+1} \leftarrow \frac{n_j^{m+1}}{n}$$
$$\mu_j^{m+1} \leftarrow \frac{1}{n_j^{m+1}} \sum_{i=1}^{n} z_{ij}^m(\theta^m) y_i$$
$$\Sigma_j^{m+1} \leftarrow \frac{1}{n_j^{m+1}} \sum_{i=1}^{n} z_{ij}^{m+1}(\theta^m)(y_i - \mu_j^{m+1})(y_i - \mu_j^{m+1})^\mathsf{T}$$
$m \leftarrow m+1$
jusqu'à $z^m = z^{m-1}$ *ou* [ou $m = mMAX$]
$\hat{z} \leftarrow z^{(m-1)}$
$\hat{\theta} \leftarrow \theta^m$

FIG. 3.7: *Un lancer de l'algorithme CEM.*

On peut voir que ces critères comme une forme de vraisemblance pénalisée, ceci dans le cadre d'une estimation des paramètres par maximum de vraisemblance [28]. Pour comprendre leur principe, on peut tout d'abord remarquer que dans un tel cadre, la vraisemblance $L(\hat{\theta}_M)$ des paramètres estimés tend à croître avec le nombre de degrés de libertés du modèles M. En effet, soit un modèle M_A dont le domaine des paramètres libres $\Omega(M_A)$ est inclus dans celui $\Omega(M_B)$ d'un autre modèle plus général M_B. Alors le vecteur de paramètre $\hat{\theta}_{M_A}$ qui maximise la vraisemblance dans le domaine $\Omega(M_A)$ appartient au domaine $\Omega(M_B)$. Sa vraisemblance $L(\hat{\theta}_{M_A})$ est donc inférieure ou égale à celle du vecteur des paramètres $L(\hat{\theta}_{M_B})$ qui maximise la vraisemblance sur le domaine $\Omega(M_B)$. En notant

$$L_{\max}(M) = L(\hat{\theta}_M) \text{ où } \hat{\theta}_M = \arg\max_{\theta \in \Omega(M)} L(\theta)$$

on a donc

$$\Omega(M_A) \subset \Omega(M_B) \Rightarrow L_{\max}(M_A) \leq L_{\max}(M_B).$$

De ce qui précède, on peut déduire que la vraisemblance croît avec le nombre de classes lorsqu'on considère des modèles de mélange à proportions variables. Ainsi, bien que la vraisemblance donne une mesure de l'ajustement du modèle aux données, elle ne permet pas de sélectionner directement le modèle le plus adéquat pour un échantillon donné.

Le principe de des critères d'information consiste à choisir le modèle qui fait croître la vraisemblance le plus possible, tout en minimisant la complexité du modèle. Pour cela, la plupart des critères se basent sur le maximum de vraisemblance pénalisé par le nombre de paramètres libres du modèle, ce qui donne l'expression généralement suivante à maximiser sur les différents modèles en compétition :

$$CI(M) = -2L_{max}(M) + \tau_{CI}Q(M)$$

La fonction $Q(M)$ indique le nombre de paramètres libres du modèle M, c'est-à-dire la dimension du domaine $\Omega(M)$. Elle représente la complexité du modèle M. Le coefficient τ_{CI} représente la pénalisation de la complexité du modèle spécifique au critère CI. Par exemple, le critère *Akaike information criterion* (AIC) proposé par Akaike [2] s'écrit :

$$AIC(M) = -2L_{max}(M) + 2Q(M) \tag{3.18}$$

on a donc $\tau_{CI} = 2$. Bozdogan [15] propose une variante du critère d'Akaike appelée AIC3:

$$AIC(M) = -2L_{max}(M) + 3Q(M) \tag{3.19}$$

donc $\tau_{CI} = 3$. Le critère *Bayes information criterion* (BIC) est obtenu par Schwarz [111] comme une approximation de la solution bayésienne exacte au problème de sélection de modèle :

$$BIC(M) = -2L_{max}(M) + \log(n)Q(M) \tag{3.20}$$

Ici, la pénalisation $\tau_{CI} = \log(n)$ fait intervenir le nombre d'observations.

Biernacki et al. dans [10] proposent d'utiliser la vraisemblance classifiante pour pénaliser le modèle gaussien et le nombre de composantes gaussiennes. Ce critère dénoté par ICL est donné par l'équation 3.21. Sur des données simulées dans la dimension deux Biernacki et al. prouvent que ICL est mieux adapté que d'autres critères lorsque la forme de ces données n'est pas gaussienne.

$$ICL(M) = -2L_M + Q_M ln(n) - 2\sum_{i=1}^{n}\sum_{j=1}^{J}\hat{c}_{ij}t_{ij}, \tag{3.21}$$

où $\hat{c}_{ij}$ représente la partition estimée déduite des probabilités a posteriori t_{ij}.

A la différence des critères d'informations cités jusqu'ici, Celeux et al. [23] proposent d'utiliser l'entropie $E(M) = -\sum_{i=1}^{n}\sum_{j=1}^{J}t_{ij}\text{log}t_{ij} \geq 0$ dans la sélection de la structure du mélange gaussien. L'entropie est vue comme une mesure de la capacité de fournir une partition adéquate des données par un mélange gaussien de J-composantes. Ce critère à minimiser est donné par :

$$NEC(M) = \frac{E(M)}{L_M - L_M^1} \tag{3.22}$$

où L_M^1 dénote la vraisemblance maximale pour une seule gaussienne.

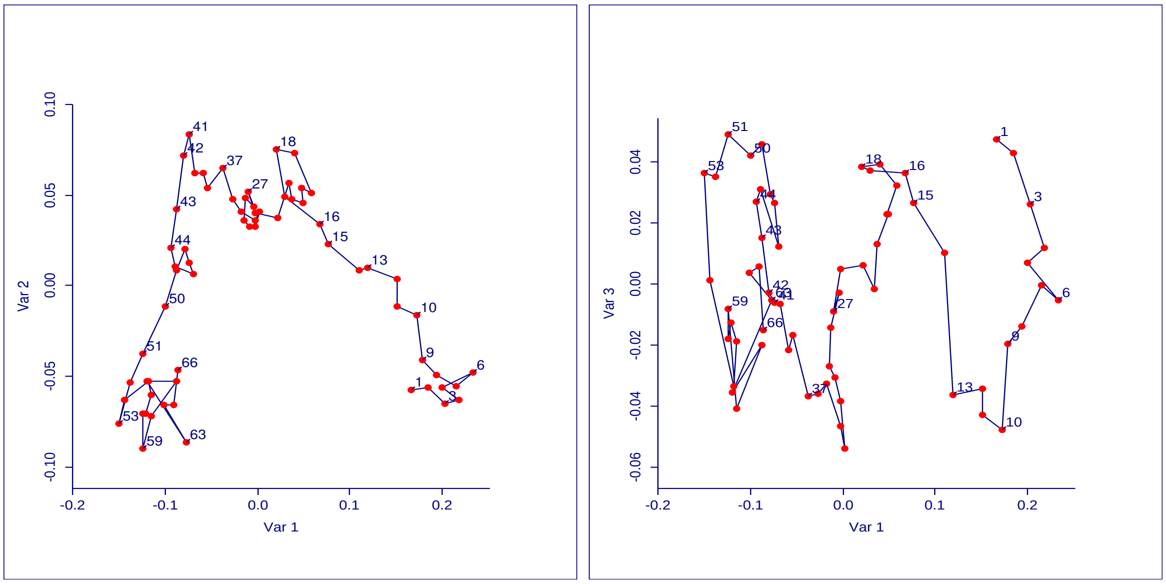
Comme nous le verrons dans le chapitre suivant (partie expérimentale et discussion), le choix d'un critère influence considérablement les résultats de classements des objets : mieux la variabilité intra-plan est modélisée, plus l'appariement et le classement des objets sont bons. Des taux de bon classements du même ordre de grandeur sont obtenus lorsque ICL ou BIC sont employés. Par contre le critère NEC fournit des mauvais résultats surtout lorsque la variabilité d'un objet suivi est modélisée par une seule gaussienne. Ce qui peut être expliqué par la normalisation de l'entropie par le terme L_M^1.

#405 #410 #415 #420

#425 #430 #435 #440

#445 #450 #455 #460

FIG. 3.8: *Sous ensemble (12/66) des apparences intra-plan de l'objet suivi "voiture Ford" de la séquence vidéo "Avengers".*



(a) (b)

FIG. 3.9: *Représentation des histogrammes de couleurs de l'objet suivi "voiture Ford" (figure 3.8) dans le premier (a) et le deuxième plan factoriel (b) de l'espace réduit à* 10 *dimensions. Un segment de droite relie deux apparences successives.*

3.7 Expérimentation

Cette section décrit une étude expérimentale de l'approche de modélisation statistique de la variabilité intra-plan des objets suivis. Dans cette démarche on procède de la manière suivante. Un objet suivi de n apparences dans le plan vidéo, est décrit par un nuage de n point dans l'espace de descripteurs "D" de dimension d. Dans l'espace de descripteurs réduit par ACP de dimension d_E selon une qualité de représentation dans cet espace de 95%, on modélise la variabilité intra-plan. Dans une **première expérience** on fixe le modèle gaussien M_i $(i = 1..7)$ et on fait varier seulement le nombre de composantes gaussiennes de 1 à une borne supérieure (MaxNbC) fixée à priori. Dans une **seconde expérience** on fixe le nombre de composantes gaussiennes de mélange gaussien et on fait varier le modèle M_i dans la famille des modèles. Dans une **troisième expérience** on fait varier ces deux éléments du mélange gaussien. On utilise dans ces expériences le critère ICL qui, d'après la discussion ci-dessus, est le critère théorique le mieux adapté pour le choix de la structure du mélange.

Pour valider cette démarche, on présente ici seulement les résultats sur l'objet suivi "voiture Ford" de 66 apparences, extrait du corpus vidéo "Avengers". Des expérimentations intensives ont été réalisées dans les deux chapitres suivants pour un objectif d'appariement et de classification d'objets. La figure 3.8 illustre quelques unes des apparences intra-plan de la "voiture Ford". On extrait de ces apparences des histogrammes de couleurs, calculés dans l'espace RGB, de 64 cellules

La figure 3.9 montre la distribution de cet objet suivi dans les deux premiers plans

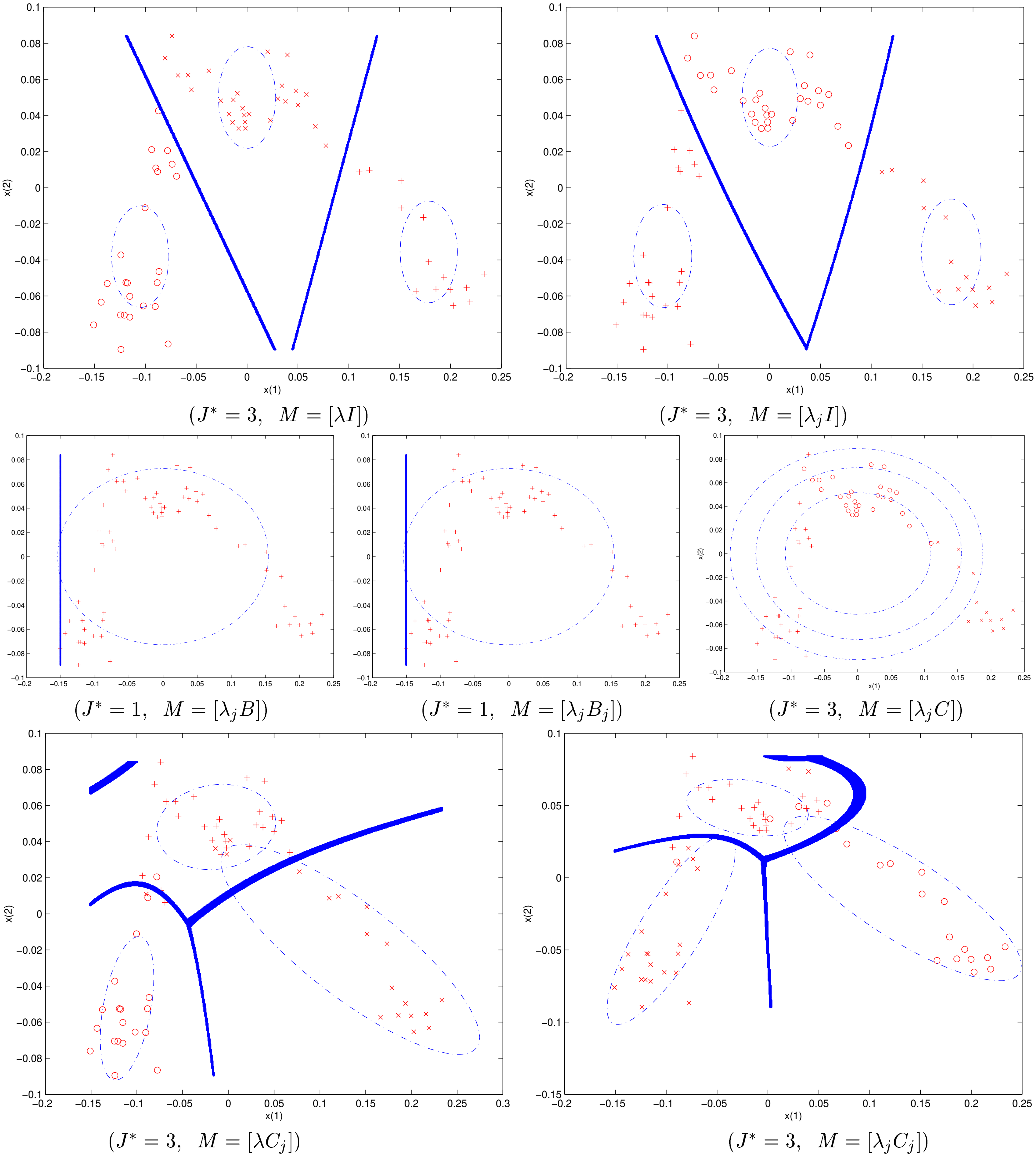
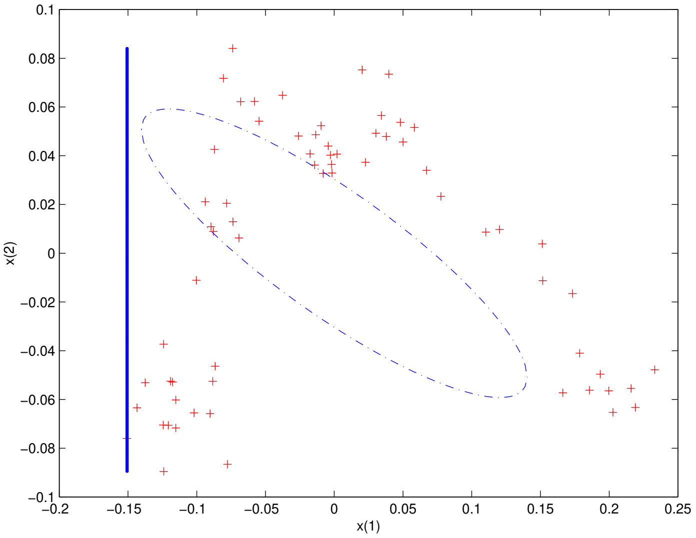
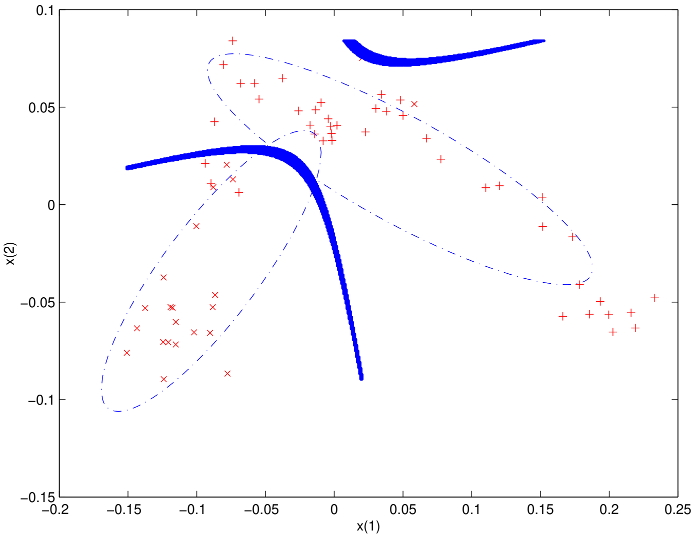


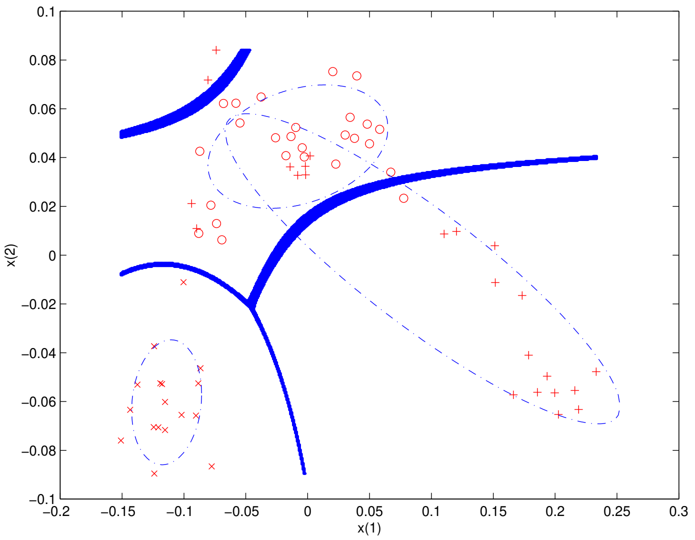
FIG. 3.10: *Illustration des résultats de modélisation des données de la figure 3.9.a, dans le cas où le modèle gaussien M est connu a priori et où seul le nombre de composantes gaussiennes est en compétition. Le nombre choisi par ICL est désigné par J^*. Pour chaque modèle la partition ainsi que les ellipses de variances et les frontières sont illustrées.*



$(J = 1, \quad M^* = [\lambda C_j])$

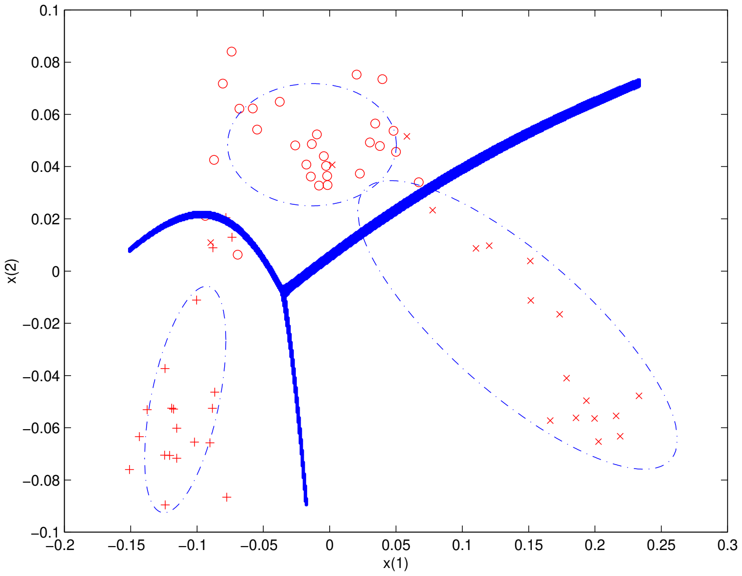


$(J = 2, \quad M^* = [\lambda C_j])$



$(J = 3, \quad M^* = [\lambda_j C_j])$

FIG. 3.11: *Illustration des résultats de modélisation des données de la figure 3.9.a dans le cas où le nombre de composantes gaussiennes J est connu a priori et où seul les modèles gaussiens sont mises en compétition. Le modèle sélectionné par ICL est désigné par M^*.*



$(J^* = 3, \quad M^* = [\lambda C_j])$

FIG. 3.12: *Illustration des résultats de modélisation des données de la figure 3.9.a dans le cas où les modèles gaussiens et le nombre de composantes gaussiennes sont mises en compétition. La structure sélectionnée par ICL est désignée par $(J^*, \quad M^*)$.*

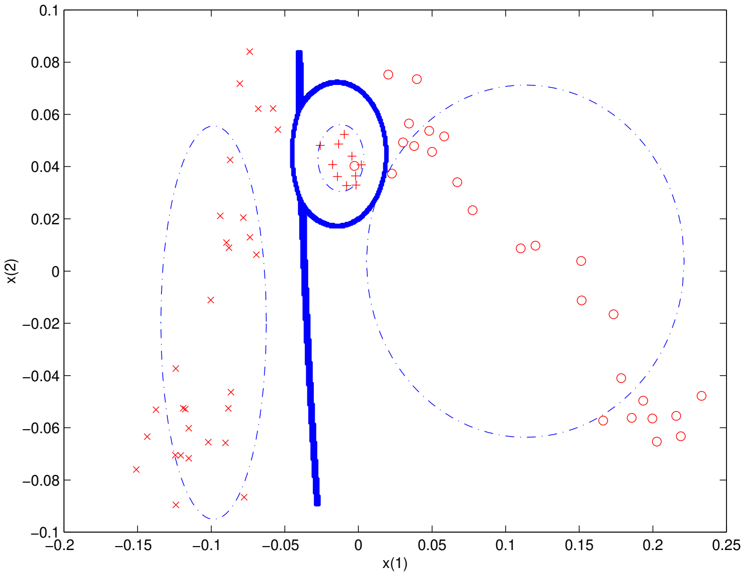
factoriels de l'espace réduit de dimension $d_E = 10$. Lors des trois expériences décrites précédemment on fixe MaxNbC à 3 pour l'exemple de cette figure. Ce nombre présente le vrai nombre de classes d'apparences intra-plan (voir figure 3.8) qu'on souhaite retrouver automatiquement par le critère ICL. D'autre part, il reste raisonnable vu le petit nombre d'apparences de l'objet modélisé.

Les figures 3.10, 3.11 et 3.12 illustrent les résultats de la modélisation lors de la première, deuxième et troisième expérience respectivement. Les partitions obtenues ainsi que les ellipses de variances et les frontières inter-classes d'apparences sont projetées dans le premier plan factoriel seulement. Rappelons que la modélisation est effectuée dans un espace de descripteurs à 10 dimensions.

Commentaires La distribution multimodale de la "voiture Ford" dans le premier plan factoriel est bien décomposable en trois classes gaussiennes.

Les modèles gaussiens avec contraintes ont été introduits dans ce travail pour palier le problème de la surestimation des paramètres gaussien du modèle général lorsque l'objet suivi dispose de peu d'apparences intra-plan par rapport à la dimension de l'espace de descripteurs. D'un autre coté, le mélange gaussien dispose d'un ensemble de critères qui permet de sélectionner la structure la plus adaptée aux données.

Les résultats de la première expérience (figure 3.11) montrent que les deux modèles gaussiens diagonaux $\lambda_j B$ et $\lambda_j B_j$ sont mal adaptés aux données. Ces modèles imposent une orientation vers les axes (figure 3.4). Ces contraintes sont difficiles à réaliser sur les



$(J^* = 3, \quad M^* = [\lambda_j B_j])$

FIG. 3.13: *Une partition alternative de celle illustrée dans la figure 3.12, obtenue en partant d'une initialisation de EM très différente.*

données modélisées. Il en résulte une mauvaise sélection du nombre de classes par le critère ICL. Pour la même raison, on trouve que le modèle $\lambda_j C$ est mal adapté à ces données car il imposent d'une même orientation et des formes ellipsoïdes égales pour les trois classes choisies par ICL. Les modèles restant fournissent des partitions proches des vraies partitions, avec des frontières linéaires et quadratiques. La frontière optimale est fournie par le modèle le plus général.

Les résultats de la deuxième expérience (figure 3.10) montrent que lorsque le nombre de classes est fixé à 2 et 3 les deux modèles $\lambda_j C$ et $\lambda_j C_j$ sont sélectionnés respectivement. Dans le cas où le nombre de classes est fixé à 2, on peut nettement voir que le modèle général est bien adapté aux données.

Dans la troisième expérience (figure 3.12) la structure du mélange gaussien semble être bien adaptée aux données "voiture Ford". Trois classes d'apparences ont été identifiées et les trois matrices de variances ont des volumes égaux mais des orientations et des formes différentes.

En effet, l'adaptation d'une structure fixe de la loi du mélange sur des données de complexité variable n'est pas la bonne stratégie à appliquer. Par ailleurs le critère comme ICL et BIC peuvent palier cet handicap. Cependant ces critères ne garantissent pas toujours la sélection de la meilleure structure de la loi de mélange. Par exemple, la figure 3.13 décrit une deuxième partition obtenue par ICL lorsque le nombre d'initialisation de l'algorithme EM est changé, ce qui conduit à une convergence assez différente. La convergence de EM et la dépendance de la solution initiale ont été déjà discutées dans la section 3.6.5.

3.8 Conclusion

Dans ce chapitre, nous avons abordé en détail le problème de la variabilité intra-plan d'un objet suivi dans l'espace de descripteurs. Cette variabilité intra-plan est dûe principalement aux changements d'apparence de l'objet à travers le temps et au manque de robustesse des descripteurs classiques à ces changements de l'image. Ces deux facteurs conduisent à une représentation spatiale dans l'espace de descripteurs qui n'est pas compacte et souvent multimodale. Une telle représentation représente un handicap majeur pour tout processus de comparaison de deux objets suivis – apparier toutes les apparences deux à deux et comment prendre une décision ? ou bien apparier juste les deux apparences représentatives des objets suivis ? –. Pour permettre une comparaison fiable de deux objets nous avons proposé ici de représenter un objet suivi par des modèles statistiques. D'une part ces modèles capturent la variabilité intra-plan et d'autre part ils ont une taille raisonnable. Il s'agit ici des modèles gaussiens multivariés qui représentent les classes d'apparences intra-plan des objets suivis. Nous avons adopté le mélange gaussien pour ses propriétés intéressantes vis-à-vis du problème étudié :

- il est suffisamment général pour modéliser des distributions complexes;
- il est suffisamment efficace pour modéliser une distribution avec peu ou beaucoup d'observations;
- il possède des critères qui permettent de sélectionner la structure du mélange gaussien (modèle gaussien et nombre de composantes gaussiennes) automatiquement.

Ce chapitre a décrit notre sélection des descripteurs de bas niveaux pour caractériser les apparences d'un objet suivi : les histogrammes de couleurs bruts et normalisés. Les expérimentations illustrées ici nous permettent de tirer les conclusions suivantes :

- la distribution d'un objet suivi dans l'espace des histogrammes de couleurs est multimodale et correspond bien à un mélange de lois gaussiennes;
- les modèles gaussiens qui imposent des contraintes sur l'orientation de la matrice de variance sont moins adaptés en pratique que les autres;
- les critères du mélange gaussien permettent un choix d'une structure de mélange gaussien qui est bien adaptée aux données modélisées mais qui n'est pas toujours optimale.

L'extension d'une telle approche sur des descripteurs locaux autour des points d'intérêts extraits des apparences est une tâche délicate car il s'agit dans un premier temps de suivre un point d'intérêt d'une apparence à une autre sachant que l'objet est non rigide; dans un second temps il faut construire une loi de mélange pour chaque point suivi, et enfin mettre en correspondance un point requête avec les n lois de mélanges des n points d'intérêts de l'objet.