UNIVERSITÉ BLIDA 1

Faculté des Sciences
Département d'Informatique

T H È S E DE DOCTORAT
TROISIÈME CYCLE

Spécialité : Ingénierie des Logiciels

INFÉRENCE BAYÉSIENNE POUR LA RECONNAISSANCE
D'ÉVÉNEMENTS DANS LES ALBUMS

Par

Siham BACHA

devant le jury composé de :

Pr Guessoum A.	U. Blida 1	Président
Pr Benblidia N.	U. Blida 1	Directrice de thèse
Pr Abed H.	U. Blida 1	Examinatrice
Pr Oulebsir Boumghar F.	USTHB	Examinatrice
Pr Ait Aoudia S.	ESI	Examinateur

Blida, Juin 2018

RESUME

d'images. Elle se réfère à leur interprétation à partir d'une perspective humaine. Dans cette dernière, la sémantique d'une image pourrait être simple ou complexe selon le type du contenu d'image. Parmi les différents types de collections d'images existantes, nous nous sommes intéressés aux albums photos numériques. Nos travaux de thèse s'inscrivent dans l'annotation sémantique d'images pour la recherche. Plus précisément, nous nous intéressons à la reconnaissance d'événement dans les albums photos personnelles.

Dans un premier temps, nous cherchons les caractéristiques qui permettent de décrire les photos de manière précise. Pour cette fin, nous utilisons les scènes et les objets comme unités sémantiques qui distinguent les événements. Outre le problème de combinaison de plusieurs caractéristiques parvenues de plusieurs images, nous devons étudier le pouvoir discriminatif de ces caractéristiques pour obtenir une reconnaissance d'événements plus précise. Afin de palier à ces problèmes, nous proposons un modèle graphique probabiliste pour la reconnaissance d'événement dans les albums photos, qui combine les scènes et les objets extraits des images d'un album, tout en introduisant la notion de pertinence de caractéristiques.

Afin de valider notre modèle, nous effectuons une étude expérimentale sur une très grande base d'albums photos appelée 'PEC'. Les résultats retournés par notre modèle sont très satisfaisants et supérieurs aux méthodes de l'état de l'art en terme de précision de classification. Ces résultats montrent l'intérêt de notre solution pour la reconnaissance d'évènements dans les albums et ouvrent de nouvelles perspectives de recherche. Par ailleurs, en se basant sur l'approche proposée dans cette thèse, il est possible de mettre en œuvre un système d'indexation d'albums de photos par reconnaissance d'évènements. Ce système permettra de faciliter la gestion intelligente et la recherche avancée des albums de photos, après avoir intégré les différents modules proposés dans cette thèse.

Mots-clés : Reconnaissance d'événements, album photos, pertinence de caractéristiques, réseau bayésien.

ABSTRACT

Automatic image annotation is a process that requires a deep analysis of the image content. It refers to their interpretation from a human perspective. In the latter, the semantics of an image could be simple or complex depending on the type of image content. Among the different types of existing images collections, we were interested in digital photo albums. Our thesis work can be classed in research axe of the semantic annotation of images for retrieval. More specifically, we are interested in event recognition in personal photo albums. Personal photos present a big challenge as their visual content is diversified.
First, we look for the characteristics that make it possible to describe the photos precisely. For this purpose, we use scenes and objects as semantic units that distinguish event categories. Once the images are characterized, we try to combine the information coming from all the images of the album. In addition to the problem of combining several features from several images, we must study the discriminative power of these features to obtain more accurate event recognition. To solve these problems, we propose a probabilistic graphic model for the recognition of event in photo albums, which combines the scenes and the objects which are extracted from the images of an album, while introducing the notion of relevance of features.
In order to validate the proposed model, we carry out an experimental study on a very large dataset of photo albums called 'PEC'. The results returned by our model are very satisfaying and superior to the methods of the state of the art in term of accuracy of classification. These results show the interest of our solution for the recognition of events in photo albums and open up new perspectives of research. Moreover, based on the approach proposed in this thesis, it is possible to implement a system of indexing photo albums by event recognition. This system will facilitate the intelligent management and advanced retrieval of photo albums after integrating the various modules proposed in this thesis.

Keywords : Event Recognition, photo album, feature relevance, bayesian network.

ملخص

شرح الصور اوتوماتيكيا عن طريق منحها و ربطها بكلمات لشرح محتواها تعتبر عملية صعبة. حيث ان هته الاخيرة اصبحت محور اهتمام العديد من البحوث العلمية في السنوات الاخيرة. تستخدم هته البحوث العديد من الوسائل مستمدة من علوم الإحصاء الذكاء الاصطناعي,....

قمنا في هته الاطروحة بدراسة مسالة التعرف على الاحداث التي تم فيها التقاط الصور. اهتممنا بحل هته المسالة في حيز الصور الشخصية و بالأخص في ألبومات الصور الرقمية. حيث ان محتوى الصور الشخصية يعد صعب التحليل بسبب تغير محتواها. لاحظنا عدم توفر بحوث مهتمة بحل هته المسالة على عكس ما هو متوفر بالنسبة للفيديو.

لحل هته المسالة قمنا بدراسة الطريقة المستخدمة من طرف الانسان لتحليل محتوى الصور الشخصية. و قمنا بتصميم حل مستوحى من مفهوم الانسان للصور ,و الذي يعتمد على وجود دلالات بسيطة أو معقدة حسب نوع محتوى الصورة. لتحقيق هذه الغاية, قمنا باستخراج الخصائص التي تصف محتوى الصور بدقة. هته الخصائص تتمثل في المشاهد والأشياء التي تعتبر الوحدات الاساسية المكونة للصورة. لكي نتمكن من تمييز الأحداث نقوم بدمج الخصائص المستخرجة من جميع صور الألبوم الرقمية. إلى جانب صعوبة مزج الخصائص المستخرجة من صور متعددة، يجب علينا أن ندرس القوة التمييزية لهته الخصائص في عملية التعرف على الأحداث. لتحقيق هذا, نقترح نموذجا يعتمد علي الاحتمالات للتعرف علي الاحداث في الألبومات الرقمية، و الذي يمكن من الجمع بين المشاهد والأشياء المستخرجة من الصور، بالإضافة الى الاخذ تعين الاعتبار القوة التمييزية لهته الخصائص.

للتحقق من صحة النموذج المقترح, قمنا باجراء دراسة تجريبية على قاعدة كبيرة جدا من الألبومات.النتائج المتحصل عليها مرضية جدا ومتفوقة على عدد من الطرق المقترحة سابقا. مما يسمح بفتح آفاق جديدة للبحث. استناداً إلى النهج المقترح في هذه الأطروحة ، يمكن تطبيق نظام لفهرسة ألبومات الصور عن طريق التعرف على الحدث. سيسهل هذا نظام الإدارة الذكية والبحث المتقدم للألبومات بعد دمج الوحدات المختلفة المقترحة في هذه الرسالة.

الكلمات المفتاحية البوم الصور, تحديد الاحداث, اهمية الخصائص, شبكة بايزية.

REMERCIEMENTS

Je tiens à remercier en premier lieu ma directrice de thèse : Pr Nadjia Benblidia d'avoir été à l'écoute de mes propositions, et de mes aspirations ainsi que pour le soutien incontestable qu'elle m'a apporté. Elle s'est investie pour que je m'approprie le sujet, tout en me laissant beaucoup de liberté quant à la démarche à adopter.

Je remercie également Dr Djamel Gaceb pour les nombreuses discussions qu'on a eu ensemble.

Il me tient à cœur de remercier la personne qui m'a ouvert le portes de son laboratoire le Pr Mohand Saïd Allili. Je lui suis très reconnaissance pour l'ensemble de connaissances qu'il m'a transmises, pour la qualité de ses explications, et pour sa disponibilité.

Je remercie très sincèrement les membres de mon jury d'avoir accepté de participer à mon évaluation. J'exprime ma reconnaissance à Madame Fatima Oulebsir Boumghar, professeur à l'université des Sciences et de la Technologie Houari Boumadiene, d'avoir accepté d'examiner ce travail. Madame Hafidha Abed, professeur à l'université de Blida 1. Une pensée à Monsieur Ait Aouadia Samy, professeur à l'Ecole Nationale d'Informatique, qui a accepté d'évaluer ce travail, paix à son âme. Merci à Monsieur Abderrezak Guessoum, professeur à l'université de Blida 1, de m'avoir fait l'honneur de présider mon jury de thèse.

Je remercie Dr Djamel Benouar, responsable de la formation doctorale en Informatique et systèmes embarqués à l'université de Blida 1, pour la bonne formation qu'il nous a assuré.

Je souhaite exprimer ma gratitude à Mustapha Ait Akkache pour les discussions constructives et fructueuses que nous avons partagées. Je le remercie de s'être plongé avec moi dans les équations pour mieux les comprendre, et d'avoir fait l'effort pédagogique de m'expliquer plusieurs subtilités qui m'échappaient avec beaucoup de passion et de bonne humeur.

Un grand merci à ma famille qui m'a apporté un soutien inconditionnel pendant cette thèse. Merci à mes parents de m'avoir transmis leur passion pour les sciences et leur encouragements durant tout mon cursus. Merci à mon père et mon professeur d'avoir été toujours là pour moi. Merci à ma sœur et mon frère de m'apporter autant de bonheur.

Je remercie mes amis qui m'entourent, et qui ont toujours cru en moi. Je pense particulièrement à Melly et à Lamia et bien d'autres.

TABLE DES MATIERES

RESUME 1

REMERCIEMENTS 4

TABLE DES MATIERES 8

INTRODUCTION 11

1 RECONNAISSANCE D'ÉVÉNEMENTS DANS LES IMAGES 16

1.1 Introduction . . . 16
1.2 La gestion et la recherche d'images . . . 16
1.2.1 Qu'est-ce qu'une image ? . . . 17
1.2.2 Digital albuming . . . 18
1.2.3 Système de recherche d'images . . . 18
1.3 Aperçu sur l'annotation sémantique d'images . . . 20
1.3.1 Définition de l'annotation sémantique d'images . . . 20
1.3.2 Fossé sémantique . . . 20
1.3.3 Annotation automatique d'images . . . 22
1.4 La reconnaissance d'événements dans les photos personnelles . 27
1.4.1 Les caractéristiques sémantiques . . . 28
1.4.2 L'événement : un concept de haut niveau . . . 30
1.4.3 État de l'art : principaux travaux . . . 30
1.4.4 Discussion et positionnement de l'approche proposée . 35
1.5 Conclusion . . . 38

2 APPRENTISSAGE AUTOMATIQUE 39

2.1 Introduction . . . 39
2.2 Apprentissage automatique . . . 39
2.3 Types d'apprentissage automatique . . . 41
2.4 Approches discriminatives . . . 43
2.4.1 Le k plus proches voisins . . . 43
2.4.2 Les machines à vecteurs de supports . . . 44

2.4.3 Les réseaux de neurones . . . 46
2.4.4 Transfer learning . . . 54
2.5 Sélection de caractéristiques pour la classification . . . 57
2.5.1 Pertinence de caractéristiques . . . 57
2.5.2 Méthode par filtre (Filter) . . . 58
2.5.3 Méthodes enveloppantes (Wrapper) . . . 58
2.5.4 Méthode intégrée (Embedded) . . . 59
2.6 Conclusion . . . 60

3 LES RÉSEAUX BAYÉSIENS 61
3.1 Introduction . . . 61
3.2 Définition des réseaux bayésiens . . . 62
3.3 Indépendance conditionnelle et d-séparation . . . 63
3.3.1 Indépendance conditionnelle . . . 63
3.3.2 d-Séparation . . . 64
3.4 Construction d'un réseau bayésien . . . 64
3.4.1 Apprentissage de structure . . . 65
3.4.2 Apprentissage des paramètres . . . 67
3.5 Inférence Bayésienne . . . 68
3.5.1 Théorème de Bayes . . . 68
3.5.2 Maximum de Vraisemblance (MV) . . . 69
3.5.3 Les lois a priori . . . 70
3.5.4 Marginalisation . . . 72
3.5.5 Théorie de la décision : Estimateur du maximum a posteriori (MAP) . . . 73
3.6 Exemples de modèles usuels . . . 74
3.6.1 Loi de Bernoulli . . . 74
3.6.2 Loi Multinomiale . . . 76
3.6.3 Loi Multinoulli . . . 76
3.7 Conclusion . . . 77

4 MODÉLISATION 78
4.1 Introduction . . . 78
4.2 Extraction des caractéristiques sémantiques d'une image . . . 78
4.3 Structure du modèle et génération de l'album photo . . . 79

4.3.1 Processus génératif d'un album 80
4.3.2 Probabilité jointe du modèle 85
4.4 Pertinence probabiliste des caractéristiques 86
4.5 Estimation des paramètres du modèle 87
4.5.1 Estimation des paramètres d'événement, d'objet et de scène 87
4.5.2 Estimation des paramètres de pertinence 89
4.6 Prédiction d'événements pour les nouveaux albums 90
4.7 Conclusion 92

5 EXPÉRIMENTATION ET ÉVALUATION 94
5.1 Introduction 94
5.2 Présentation des bases de tests 94
5.2.1 Personal Event Collections (PEC) dataset 94
5.2.2 La base ImageNet 97
5.2.3 La base Places205 97
5.3 Critères d'évaluation 98
5.4 Représentation d'image 100
5.5 Évaluation des modules scène/objet de réseau bayésien 101
5.6 Évaluation de la pertinence des caractéristiques 104
5.7 Comparaison de notre méthode avec les travaux existants . . . 108
5.8 Temps de calcul 111
5.9 Conclusion 112

CONCLUSION 113

REFERENCES 116

LISTE DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

Figure 1.1 Première photographie prise par Joseph Nicéphore. . . 18
Figure 1.2 Les différents fossés existants. 23
Figure 1.3 Exemples de segmentation d'images en blocs. 26
Figure 1.4 Exemples de segmentation d'images en régions. 26
Figure 1.5 Les catégories d'images présentent dans la base COREL. 27
Figure 1.6 Exemple d'images qui peuvent appartenir à deux classes différentes 30

Figure 2.1 Méthode 4-ppv. 44
Figure 2.2 SVM binaire. 45
Figure 2.3 Méthode une-contre-reste. 47
Figure 2.4 Méthode une-contre-une. 47
Figure 2.5 Exemple d'un perceptron multi-couches. 48
Figure 2.6 Principe d'une couche de convolution dans un CNN. . . 50
Figure 2.7 Architecture du réseau de neurones à convolutions LeNet. 51
Figure 2.8 Résultats obtenus sur la base du challenge ImageNet LSVRC(ImageNet Large Scale Visual Recognition Challenge [1]. 52
Figure 2.9 bloc Inception. 53
Figure 2.10 Réutilisation d'un réseau de neurones pré-entraîné. . . 54
Figure 2.11 Utilisation d'un réseau de neurones convolutif pré-entraîné pour l'extraction des caractéristiques [2]. 55
Figure 2.12 Procédure de la méthode Filter 58
Figure 2.13 Procédure de la méthode Wrapper. 59

Figure 3.1 Connexion en série/ divergence/ convergente 63
Figure 3.2 Exemple de d-séparation. 64

Figure 4.1 Représentation graphique du modèle proposé 80
Figure 4.2 Exemples de thèmes scène latents et thèmes objet latents présents dans l'évènement "voyage". 81
Figure 4.3 Représentation graphique du modèle proposé basée uniquement sur les scènes. 83
Figure 4.4 Représentation graphique du modèle proposé basée uniquement sur les objets. 84

Figure 5.1 Exemples d'images de la base PEC. 97
Figure 5.2 Exemples d'images de la base ImageNet. 98
Figure 5.3 Exemples de scènes de la base Places205. 99
Figure 5.4 Matrice de confusion pour reconnaissance d'événements basés uniquement sur les objets (O+PGM). 101
Figure 5.5 Matrice de confusion pour la reconnaissance d'événements en utilisant uniquement les scènes pour la reconnaissance d'évènements (S+PGM). 102
Figure 5.6 Performances de détection de l'objet *"bateau"* en terme de précision et de rappel. 103
Figure 5.7 Impact du nombre d'images représentatives sur la précision de la prédiction des classes d'événements. 104
Figure 5.8 Exemple d'un événement 'd'anniversaire" tiré de la base PEC illustrant le comportement sporadique de prise de photos d'un utilisateur. 105
Figure 5.9 Variabilité du contenu d'images dans un événement "exposition". 106
Figure 5.10 Exemple d'un événement "mariage" mal classé. 106
Figure 5.11 Matrice de confusion des 14 événements en combinant les scènes et les objets (OS+PGM), sans utiliser la pertinence de caractéristiques. 108
Figure 5.12 Matrice de confusion des 14 événements en combinant les scènes et les objets, en utilisant la pertinence de caractéristiques (R+OS+PGM). 109
Figure 5.13 Temps de prédiction d'événements en fonction des tailles des albums. 112
Tableau 1.1 Les événements et leurs caractéristiques sémantiques. 31
Tableau 1.2 Étude comparative entre les approches d'annotation de photos personnelles pour la reconnaissance d'événements. . 36
Tableau 3.1 Exemples de distributions a priori conjuguées 71
Tableau 5.1 Nombre d'albums, d'images d'apprentissage et de test par classe constituant la base PEC 96
Tableau 5.2 Comparaison de notre méthode avec quelques travaux existants sur la base PEC. 110

INTRODUCTION

Contexte et problématique

Avec l'évolution technologique, on assiste à une explosion des photos numériques. Cette explosion est due au faible coût des dispositifs d'acquisition d'images (caméras numériques, téléphones mobiles, etc.) et des appareils de stockage (disques durs, ordinateurs, etc.). De plus, l'émergence des plateformes de partage d'images en ligne et des réseaux sociaux (Facebook, Flicker,...) a modifié les comportements sociaux des utilisateurs. Les utilisateurs créent et partagent de plus en plus de photos. Face à l'abondance de ces données, il devient très difficile à un utilisateur de gérer et de retrouver rapidement une information souhaitée. Diverses méthodes ont été mises en place, par les utilisateurs, en vue de faciliter l'accès aux données recherchées. Le regroupement de photos en albums est la manière la plus utilisée pour organiser les photos. Les utilisateurs ont tendances à attribuer, manuellement, aux albums des étiquettes correspondant aux événements où les photos ont été prises. L'événement (tel que anniversaire, mariage,...) représente un index sémantique important pour faciliter la gestion et la recherche des albums photos.

L'explosion du volume d'albums photos a rendu cette annotation manuelle un travail laborieux. Ce dernier est très coûteux en termes de temps et d'efforts. Par conséquent, il apparaît nécessaire de disposer de systèmes capables d'analyser, de gérer et de retrouver les albums photos automatiquement en leur attribuant des étiquettes, et donc d'avoir un système d'annotation automatique de photos. Un tel système est capable d'analyser et de déterminer le contenu sémantique d'images.

Plusieurs efforts ont été déployés pour attribuer automatiquement un événement à une ou plusieurs image(s). Ces systèmes visent à résoudre le problème de reconnaissance d'événements dans les images. Dans la littérature, on distingue deux approches pour la reconnaissance d'événements dans les photos personnelles : la reconnaissance d'événements basée sur le contenu et la reconnaissance d'événements basée sur le contexte. La première fait référence aux caractéristiques de bas niveaux des images (par exemple, texture, forme...), tandis que la deuxième fait référence aux caractéristiques du contexte externe d'image (par exemple, heure de prise de photo, localisation...).

Bien que ces approches produisent des résultats satisfaisants pour la reconnaissance d'événements dans les photos, elles s'avèrent être insuffisantes pour répondre aux besoin des utilisateurs. Cela est dû principalement au problème du fossé sémantique. Le fossé sémantique séparent les concepts sémantiques interprétables par les être humains et les caractéristiques numériques de bas-niveau. D'autre part, des images différentes peuvent avoir des signatures visuelles similaires mais des significations sémantiques différentes. Par ailleurs, les approches basées sur le contexte souffrent du problème de fossé sensoriel.

Franchir les fossé sémantique et sensoriel est une problématique de recherche qui a suscité l'intérêt de la communauté de vision par ordinateur. Pour combler ces fossés, des travaux plus récents amènent de la sémantique aux images en s'appuyant sur la représentation des photos personnelles par des caractéristiques de haut-niveau (sémantiques). De plus, l'événement reste un concept complexe qui se base lui même sur d'autres concepts sémantique. La représentation d'image par des caractéristiques sémantiques permet d'enrichir la sémantique associée à l'image afin de mieux classer l'événement.

Cette thèse s'inscrit le cadre général de l'annotation automatique d'images et plus précisément dans la reconnaissance d'événements dans les albums de photos personnelles.

Principales contributions

Notre objectif dans cette thèse est d'annoter automatiquement des albums de photos personnelles par les événements dans lesquels les photos ont été prises. En d'autres termes, trouver à quelle catégorie d'événement appartient un ensemble de photos donné. Pour ce faire, nos contributions portent sur quatre volets :

- **L'annotation sémantique d'un album photos**
  Dans notre travail, un album photos est constitué d'un ensemble de photos qui représente un événement particulier. Comme nous allons voir dans l'état de l'art établi dans le chapitre 1, plusieurs systèmes de reconnaissance d'événements ont été conçu pour annoter une seule image. Nous proposons un réseau bayésien qui prend en charge un album photos à la fois. L'analyse d'un ensemble d'images donne une idée plus com-

plète sur l'événement en question par rapport à une image individuelle. Chaque image dans l'album apporte une information sur l'événement, ce qui permet une classification plus précise.

- **La représentation des albums photos par des caractéristiques de haut-niveau**
  Nous avons examiné plusieurs caractéristiques d'images qui pourront influencer les performances de la prédiction d'événements dans les albums photos. Le fait que l'image est constituée d'objet et leur contexte visuel, nous avons conclu que les caractéristiques de haut-niveau, scène et objet, sont les plus représentatives pour les albums de photos personnelles. Dans notre travail, un album est représenté par l'ensemble de scènes et objets extraits de toutes ses images.

- **L'introduction de la notion de pertinence de caractéristique**
  Nous proposons d'intégrer la notion de pertinence de caractéristiques dans le réseau bayésien afin de prendre en considération le pouvoir discriminatif de ces caractéristiques dans la catégorisation. La pertinence estime ce qu'une caractéristique apporte pour prendre la bonne décision de catégorisation sur une classe d'événement cible. Elle permet d'identifier les caractéristiques qui discriminent le mieux une catégorie d'événement.

- **Modélisation d'un réseau bayésien pour la reconnaissance d'événement dans les albums photos**
  Nous proposons un réseau bayésien qui combine les objets et les scènes, et qui prend en considération leur pouvoir discriminatif dans la reconnaissance d'événements dans les albums photos. Le choix des réseau bayésien est motivé par deux aspects principaux. Tout d'abord, les bons résultats obtenus par l'utilisation des modèles graphiques probabiliste dans le domaine de reconnaissance d'évènements sportifs dans les images [3]. La deuxième motivation est l'avènement des méthodes bayésiennes qui offre un outil attrayant pour étudier la modélisation de la perception humaine des évènements tel que évoqué par Rand [4].

L'évaluation de nos contributions forme une partie prépondérante de ce travail. Les performances de l'approche proposée sont mises en évidence à

travers une comparaison avec d'autres approches de la littérature du domaine.

Organisation de la thèse

Outre l'introduction générale et la conclusion générale, la thèse est structurée en cinq chapitres. Les trois premiers chapitres sont dédiés à l'état de l'art afin de mieux situer le contexte de notre travail et les concepts sur lesquels reposent nos travaux. Les deux derniers chapitres sont consacrés à la présentation de nos contributions.

Le premier chapitre expose un aperçu général sur les concepts de base de l'annotation d'images. Nous introduisons en premier lieu la notion d'image et d'album puis nous exposons le processus général de la recherche d'images. Par la suite, nous détaillons l'annotation sémantique d'images et ses différentes catégories. Nous nous focalisons en particulier sur l'annotation automatique d'images et ses types. Finalement, nous présentons un état de l'art sur la reconnaissance d'événement dans les images et nous faisons un tour d'horizon des travaux apparentés au nôtre. Une synthèse sur les travaux existants est établie dans ce chapitre.

Le deuxième chapitre introduit les concepts fondamentaux des algorithmes d'apprentissage. Nous présentons quelques méthodes d'apprentissage discriminatives : le k plus proches voisins, les machines à vecteurs de supports et les réseaux de neurones. Nous mettons également l'accent sur les réseaux de neurones convolutifs qui trouvent de plus en plus de succès dans le domaine de la reconnaissance visuelle. Nous parcourons aussi les différentes méthodes de sélection de caractéristiques.

Dans le troisième chapitre, nous nous intéressons aux réseaux bayésiens leurs propriétés, notamment la représentation de l'indépendance conditionnelle et l'inférence bayésienne. Nous présentons aussi les distributions probabilistes les plus connues.

Le quatrième chapitre, expose notre réseau bayésien intégrant la notion de pertinence des caractéristiques pour la reconnaissance d'événements dans les albums photos. Nous présentons la structure notre modèle et comment l'utiliser pour générer un album photos. Nous présentons l'estimation des paramètres du modèle et les étapes d'inférence bayésienne pour prédire l'événe-

ment d'un nouvel album.

Le cinquième chapitre présente les expérimentations, menées sur la base d'albums photos PEC, afin de valider nos propositions. Nous établissons une étude comparative entre notre approche et d'autres approches de l'état de l'art tout en discutant les résultats obtenus. Les performances de notre modèle en terme de calcul sont aussi mises en évidence.

Enfin, la conclusion souligne l'avantage de l'approche proposée, en présentant une synthèse des contributions apportés et en évoquant les pistes définissant des perspectives possibles pour des travaux futurs.

Chapitre 1
RECONNAISSANCE D'ÉVÉNEMENTS DANS LES IMAGES

1.1 Introduction

Avec le développement des appareils photo numériques et les téléphones mobiles, la quantité des photos personnelles est devenue de plus en plus volumineuse. Pour faciliter la gestion et la recherche d'images, il est nécessaire d'annoter les images par des mots-clés, comme nous allons voir dans ce chapitre. Les albums de photos personnelles présentent une masse importante d'informations qui reste faiblement étudiée par les systèmes d'annotation d'images actuels. En effet, la plupart des systèmes d'annotation d'images actuels fournissent une sémantique simple. Or, le contenu des photos personnelles est très complexe. Le traitement de ce type d'images représente un vrai challenge pour les systèmes d'annotation.

Dans ce chapitre, nous donnons en premier lieu un aperçu sur le domaine de la recherche d'images. En second lieu, nous ciblons les rappels sur l'annotation automatique d'images, qui sont nécessaire à la compréhension de l'étude menée. En troisième lieu, nous focalisons notre étude sur la reconnaissance d'événements dans les albums photos. Nous présentons les caractéristiques sémantiques, objet et scène, utilisées pour la reconnaissance d'événements. Enfin, nous présentons les principaux travaux de la reconnaissance d'événements dans les images. Ces travaux englobent plusieurs problématiques sur lesquelles s'articule notre étude comparative. Enfin, nous terminons ce chapitre par le positionnement de notre travail par rapport à l'existant.

1.2 La gestion et la recherche d'images

Dans ce qui suit, nous allons présenter brièvement la définition d'une image, le digital albuming et les différentes techniques utilisées pour l'annotation d'images.

1.2.1 Qu'est-ce qu'une image ?

L'image est une représentation visuelle des différents sujets entre autres : objet, scène ou personne par la photographie, le dessin, etc. La première photographie (ou image) a été prise par Joseph Nicéphore Niépce en 1827 [5]. La photographie, intitulée le Point de vue de Gras, représente la vue depuis une fenêtre de sa propriété de Saint-Loup-de-Varennes, sur une plaque d'étain recouverte de bitume de Judée. Elle est illustrée dans la figure 1.1. En 1975, l'ingénieur américain Steven Sason [6] a créé le premier appareil photo électronique avec lequel la première image numérique a été prise. Les images numériques sont des images qui peuvent être stockées sur des supports informatiques tels que le téléphone, ordinateur, etc.

Avec la révolution numérique au cours de ces dernières décennies, la quantité des images a connu une croissance significative. Elles sont présentes dans plusieurs domaines notamment médical, tourisme, culturel, etc. Selon le fondateur de Facebook [1], toutes les secondes plus de 2263 photos sont mises en ligne sur Facebook, soit 2716000 photos toutes les 20 minutes et plus de 71.425 milliards photos par an. Par ailleurs, d'après les statistiques réalisées par le site Brandwatch [2] en 2016, plus de 80 millions de photos sont chargées chaque jour sur le site Instagram, un site Web qui permet aux gens de partager leurs images.

Deloitte Global [3] estime qu'en 2016, 2.5 trillions de photos ont été partagées ou stockées en ligne, soit une augmentation de 15% par rapport à l'année 2015. Environ trois quarts de ce total seront probablement partagés, et le reste sera sauvegardé en ligne. Toutes ces photos sont stockées sur les disques durs des ordinateurs,ou sur des plateformes de partage de photo telles que Picasa [4] ou Flickr [5], et elles présentent une quantité énorme d'informations à gérer. A cause de l'accroissement constant en nombre d'images, l'organisation et la

1. http ://www.planetoscope.com/Internet-/1217-nombre-de-photos-deposees-sur-facebook.html, dernière visite : mai 2018. Le site donne mis à jours le nombre de photos chargées sur facebook en temps réel.
2. https ://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts-for-2016
3. https ://www2.deloitte.com/global/en/pages/technology-media-and-telecommunications/articles/tmt-pred16-telecomm-photo-sharing-trillions-and-rising.html
4. https://picasa.google.com/ est un site spécialisé dans le partage de photos géo-référencées.
5. http://flickr.com/ est un site généraliste de partage de photos

FIGURE 1.1 – Première photographie prise par Joseph Nicéphore [7].

recherche d'images ont été qualifiées de big challenge [8].

1.2.2 Digital albuming

Dans leur étude sur le regroupement d'événements et la qualité des albums numériques, les auteurs Loui et Savakis [9] définissent le processus de création d'album numérique (en anglais digital albuming) comme : "un ensemble de processus de regroupement d'images en événements et sous-événements. La génération d'un album se fait après avoir éliminé les images de mauvaise qualité et les images en double". Autrement dit, le digital albuming permet de créer une mise en page pour présenter un groupe de photos en fonction des photos elles-mêmes. Le résultat de ce processus est semblable à un petit album photos physique représentant un événement spécifique tel que la naissance d'un enfant ou la fête de mariage d'un ami. Dans notre travail, nous définissons un album photo par un ensemble d'images, non ordonnées dans le temps, prises dans un seul événement.

1.2.3 Système de recherche d'images

L'objectif principal d'un système de recherche d'images est de répondre aux besoins d'un utilisateur en sélectionnant les images les plus pertinentes

par rapport à ses besoins. La notion de pertinence représente la base de la définition d'un système de recherche d'image, c'est-à-dire l'adéquation entre la requête de l'utilisateur et le contenu effectif des images. Les critères de recherche dans une base d'images dépendent du domaine d'application où chaque utilisateur a ses propres critères de pertinence, en fonction de ses besoins. Un système de recherche d'images est défini comme [10] "un système pour la navigation, la récupération et la recherche d'images à partir d'une grande base de données d'images numériques". Le calcul de la pertinence d'une image pour une requête d'utilisateur comprend généralement les étapes suivantes [10] :

1. La représentation de l'image et des besoins d'utilisateur dans un langage commun, qui correspondent aux phases de formulation de requêtes et d'indexation, respectivement. Le but dans cette étape consiste à créer une signature ou un index d'image, qui sert comme caractéristique pour identifier l'image et la comparer avec d'autres.

2. La comparaison des requêtes et des images en utilisant une mesure de similitude afin de présenter à l'utilisateur les images les plus similaires qui correspondent à ses besoins.

3. L'utilisation des algorithmes de recherche qui permettent, en se basant sur les deux étapes précédentes, de retrouver rapidement les images recherchées.

4. L'utilisation d'une interface utilisateur qui facilite la formulation des requêtes et la procédure de recherche d'images.

La première étape correspond à la phase d'indexation qui détermine le mode d'interaction entre le système de recherche d'image et l'utilisateur. Autrement dit, l'indexation d'une base d'images permet d'extraire l'ensemble des caractéristiques à partir desquelles l'adéquation aux requêtes de l'utilisateur pourra être calculée. Deux types de caractéristiques peuvent être utilisés pour indexer les images [11] : les caractéristiques de haut niveau (dites sémantiques) et/ou les caractéristiques de bas niveau (dites visuelles). Selon le type d'indexation, la requête peut être formulée par l'utilisateur en fonction des descripteurs de

bas niveau ou des descripteurs sémantiques. Dans cette thèse, nous nous intéressons à l'annotation automatique des images par des événements. Cette dernière associe aux images des mots-clés qui seront utilisées ultérieurement comme index pour effectuer la recherche.

1.3 Aperçu sur l'annotation sémantique d'images

L'annotation sémantique d'images est considérée comme un des défis des ces dernières décennies. Elle a été introduite afin de combler le fossé sémantique. Nous allons présenter dans ce qui suit, sa définition et le problème du fossé sémantique.

1.3.1 Définition de l'annotation sémantique d'images

L'annotation sémantique d'images est le processus par lequel on affecte à une image des informations textuelles. Elle consiste à assigner des étiquettes à une image afin de décrire son contenu sémantique par des mots-clés. Ces derniers peuvent correspondre à tout ce qui explique le contenu de l'image à savoir : les objets, les scènes, les événements qu'on y rencontre, les personnes, les lieux, etc. L'annotation sémantique est utilisée essentiellement pour la gestion des bases d'images lors de la recherche d'images. Les images annotées sont plus faciles à retrouver que les images non annotées lors d'une recherche basée sur les mots-clés dans une base de données d'images. Elle permet une compréhension du contenu qui sera exploitable par des traitements ultérieurs. Alors que l'être humain utilise la langue naturelle pour transmettre des connaissances en utilisant des textes, l'annotation sémantique a pour vocation d'accéder au moins à une partie de ces connaissances en se basant sur une catégorisation prédéfinie à défaut de pouvoir comprendre l'image entièrement.

1.3.2 Fossé sémantique

Les images sont stockées sur les supports informatiques, sous format numérique, où la représentation de l'information est binaire. Ce format n'est pas interprétable par les êtres humains. Afin d'effectuer une recherche, l'utilisateur emploie des textes qui lui permettent d'associer un sens aux données recherchées. Nous distinguons trois types de caractéristiques : les caractéristiques

contextuelles, les caractéristiques perceptuelles et les caractéristiques conceptuelles de l'image. Nous admettons communément trois niveaux d'indexation :

1. L'analyse contextuelle, qui fait référence au contexte de l'image. Le contexte est un concept général. Il est donc difficile de lui donner une définition claire et définitive. Certains types de contexte sont, dans la pratique, plus important que d'autres. Selon Dey [12], ces types sont : la localisation où l'image a été prise, le temps d'acquisition, l'identité et l'activité des personnes apparentes dans l'image.

2. L'analyse numérique, qui fait référence aux caractéristiques primaires ou bas niveaux extraites automatiquement. Ces caractéristiques correspondent à la couleur, la texture et la forme. Elles peuvent représenter la totalité de l'image ou des régions de celle-ci.

3. L'analyse sémantique, qui fait référence au contenu interprétable de l'image représentant dans l'ensemble des objets, scènes et significations que l'on peut y associer.

Smeulders et al. [13] distinguent deux principaux problèmes liés à ces analyses : le *fossé sensoriel* (sensory gap) et le *fossé sémantique* (semantic gap). La figure 1.2 illustre un schéma explicatif de ces différents problèmes en les comparant avec la vision humaine. Le *fossé sensoriel* [13] résulte de la perte et /ou la déformation d'information lors de l'acquisition d'une image. Il y a une importante perte d'information résultant de la projection du monde réel en 3D sur une image 2D. De plus, la perte et/ou la déformation d'information peut être induite par le matériel de prise de vue (appareil photo numérique, capteur,...etc). Plusieurs problèmes peuvent être classés sous appellation de gap sensoriel : le bruit numérique, les approximations colorimétriques,...etc.

D'autre part, il existe un décalage important entre la description de bas-niveau et la description de haut-niveau de l'image. Ce décalage est appelé le *fossé sémantique*. Ce problème a été évoqué la première fois par Smeulders et al. [13], qui l'a définit par : " Le fossé sémantique est le manque de concordance entre les informations que la machine peut extraire d'un document numérique et des interprétations humaines ". Ayache [14], quant à lui, donne une autre définition :" Le fossé sémantique est ce qui sépare les représentations brutes (tableaux de nombres) et sémantiques (concepts et relations) d'un document

numérique ". L'utilisateur rencontre des difficultés dans la formulation de ses besoins, pour une éventuelle recherche, sur des interfaces où le recours au texte n'est pas un critère fonctionnel.

Récemment, plusieurs travaux ont été proposés pour pallier au problème de fossé sémantique dans les images [16, 17]. Ces approches se basent sur les avancées scientifiques dans le domaine de l'apprentissage automatique pour l'annotation sémantique d'images. L'annotation automatique d'images fournit un index efficace qui peut être utilisé pour la communication entre les utilisateurs et les systèmes de recherches d'images. Le principal avantage de l'annotation automatique est d'être objective par rapport à l'annotation manuelle puisqu'elle est effectuée par la machine. Elle permet de réduire le travail laborieux et d'économiser du temps. Dans notre travail, nous nous intéressons à l'annotation automatique d'images pour laquelle une section détaillée a été dédiée.

1.3.3 Annotation automatique d'images

Nous avons choisi de présenter l'état de l'art des travaux sur l'annotation automatique d'images (AIA) en fonction des caractéristiques utilisées pour annoter celles-ci. Ainsi, nous avons identifié dans la littérature deux approches d'annotation : approches basées sur les caractéristiques contextuelles et approches basées sur les caractéristiques visuelles.

1.3.3.1 Approches basées sur les caractéristiques contextuelle

Le contexte peut être utilisé dans différents domaines. Il est donc difficile de donner une définition claire et définitive. Il y a certains types de contexte qui, en pratique, sont plus importants que d'autres. Selon Dey [12], ces types sont la localisation, l'identité de la personne, l'activité et le temps d'acquisition. La plupart des dispositifs mobiles et des appareils photos récents sont dotés de capteurs sensoriels. Leurs capteurs intégrés, comme l'accéléromètre ou le GPS, peuvent aider à fournir des informations contextuelles utiles sur la situation ou l'événement où les images ont été prises, ce qui permet d'enrichir l'annotation. Cette nouvelle source riche d'informations redirige les efforts des chercheurs pour exploiter l'information contextuelle dans AIA. Parmi toutes les informations contextuelles disponibles, la localisation et le temps d'acquisi-

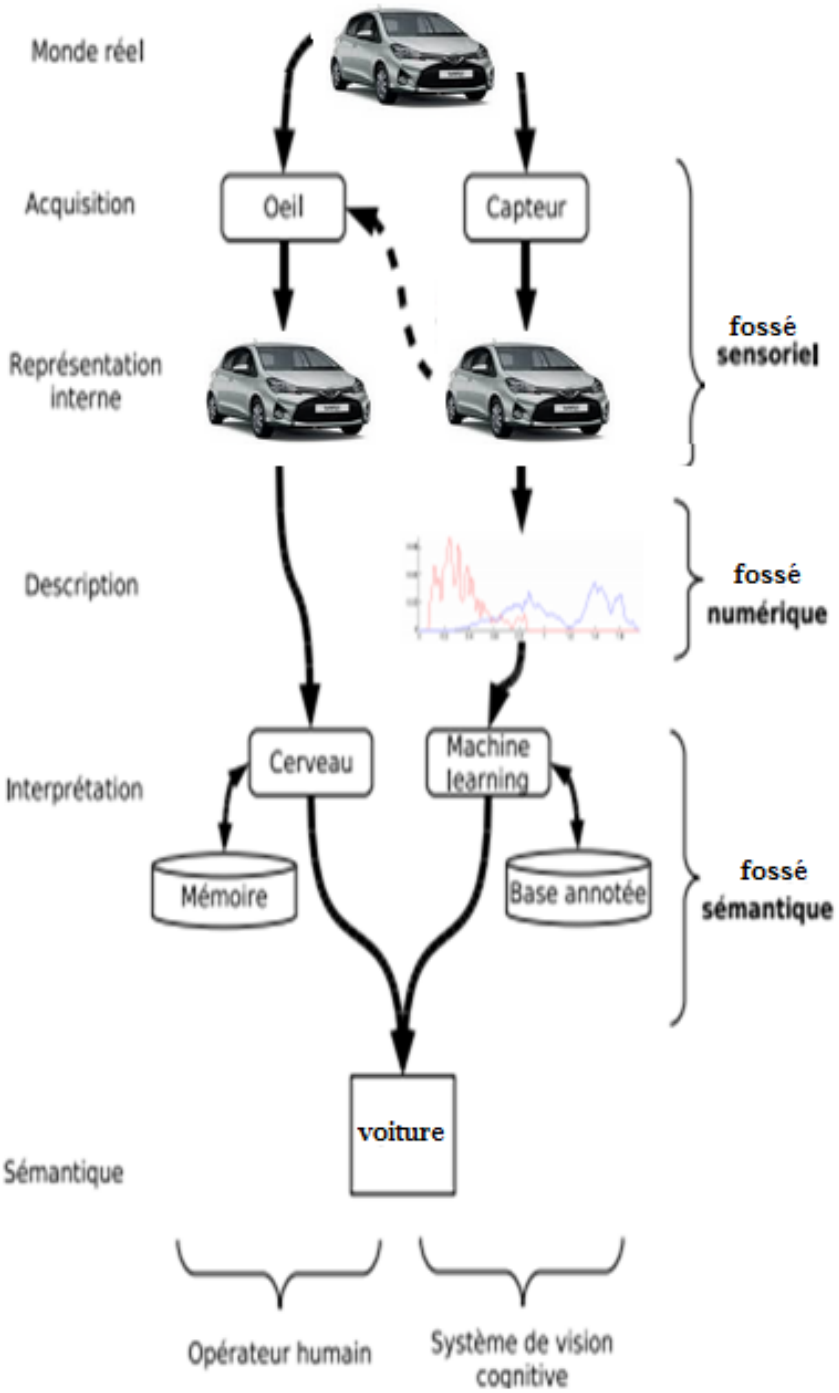

FIGURE 1.2 – Les différents fossés existants [15].

tion d'image (information spatio-temporelle) ont suscité un intérêt particulier

dans ce domaine de recherche.

Deux principaux types d'annotations d'images contextuelles sont distingués : l'approche basée sur les descripteurs de bas niveau et l'approche basée sur les descripteurs de haut niveau [18].

1. L'approche basée sur les descripteurs de bas-niveau peut fournir des annotations précises telles que : les informations EXIF, les coordonnées GPS, le temps d'acquisition ou toute autre information contextuelle qui peut être attachée comme métadonnées à l'image [18]. Cependant, les descripteurs de bas-niveau ne sont pas toujours faciles à interpréter par l'utilisateur, par exemple les coordonnées GPS 36.752872, 3.042027. Ainsi, le fossé sémantique entre la représentation de l'information contextuelle de bas niveau et la représentation de haut niveau de la même information n'est toujours pas comblé.

2. L'approche basée sur des descripteurs de haut-niveau fournit des sémantiques et des informations sur le contexte qui sont facilement interprétables par l'utilisateur [19, 20]. Par exemple, les coordonnées GPS peuvent être utilisées facilement par l'utilisateur lorsqu'ils sont liés à un nom d'une ville (Alger, coordonnées GPS : 36.752872, 3.042027).

Récemment, plusieurs approches ont été proposées dans la littérature pour annoter les images automatiquement en se basant sur les informations contextuelles de l'image. Les informations contextuelles sont faciles à extraire et aident à créer de multiples index dans le but de construire une indexation d'images efficace. Toutefois, l'utilisation des informations contextuelles présente également certaines faiblesses dont la principale étant de dégrader considérablement la précision de recherche. En effet, une erreur peut se produire lors de l'acquisition des informations contextuelles par les capteurs des appareils. Par conséquent, les systèmes d'annotation d'images basés sur le contexte peuvent générer des annotations moins précises en raison d'une mauvaise interprétation de ces informations manquantes ou détériorées. Bien que l'utilisation des informations contextuelles ait montré son utilité, elle est souvent délaissée [18].

1.3.3.2 Approches basées sur les caractéristiques visuelles

Les approches basées sur les caractéristiques visuelles peuvent être classées comme suit :

1. **Approches basées sur les caractéristiques globales**
   Les descripteurs globales sont utilisés pour caractériser une image de manière générique. Dans le domaine du CBIR (Content-Cased image Retrieval- Recherche d'image par le Contenu), ces caractéristiques globales sont calculées sur toute l'image par exemple en construisant des histogrammes ou en calculant des valeur moyennes. Plusieurs méthodes d'annotation d'images basées sur les caractéristiques globales ont été proposées dans la littérature. Par exemple, Vapnik et al. [21] utilisent les HSV (Hue Saturation Value) histogrammes de couleur pour caractériser les images, et les annoter par l'une des sept catégories définies de la base Corel en utilisant les Machines à Vecteurs de Support (SVMs) comme classificateur. Avec le même objectif, Makadia et al. [22] utilisent les caractéristiques de la couleur et la texture pour identifier les images les plus similaires. L'annotation est obtenue par l'algorithme des k plus proches voisins.

2. **Approches basées sur les caractéristiques locales**
   Dans ces approches, les images sont divisées en blocs de taille fixe ou en régions comme illustré dans la figure 1.3. Ces régions ou blocs sont appelés aussi sous-unités d'image. Les caractéristiques visuelles sont alors calculées individuellement pour chaque sous-unité et appelées descripteurs locaux.

   La deuxième méthode pour calculer les caractéristiques locales consiste à diviser l'image en région, objet ou contour homogènes en utilisant des algorithmes de segmentation comme illustré dans la figure 1.4.

3. **Approches hybrides** Les approches hybrides consistent à combiner des approches locales et globales pour annoter les images. Des chercheurs ont montré que cette combinaison peut améliorer l'annotation automatique d'images pour certaines tâches, telles que la reconnaissance de visage [24, 25, 26, 27] et la reconnaissance d'objet [28].

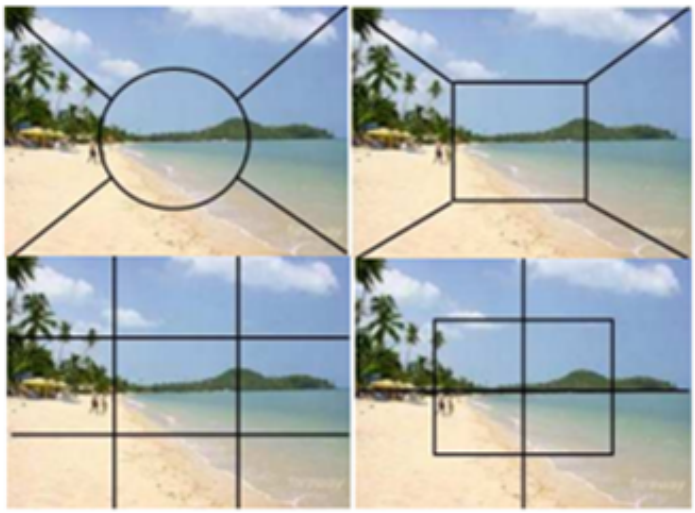

FIGURE 1.3 – Exemples de segmentation d'images en blocs [23].

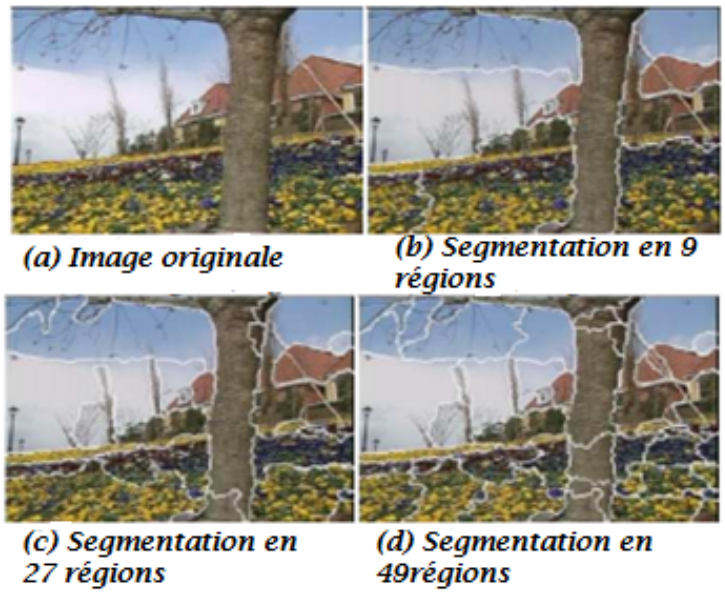

FIGURE 1.4 – Exemples de segmentation d'images en régions [23].

Une synthèse de ces approches d'annotation est présentée dans les travaux de [23]. Les approches utilisant ces différentes représentations d'image ont connu beaucoup de succès dans les bases d'images classiques telle que

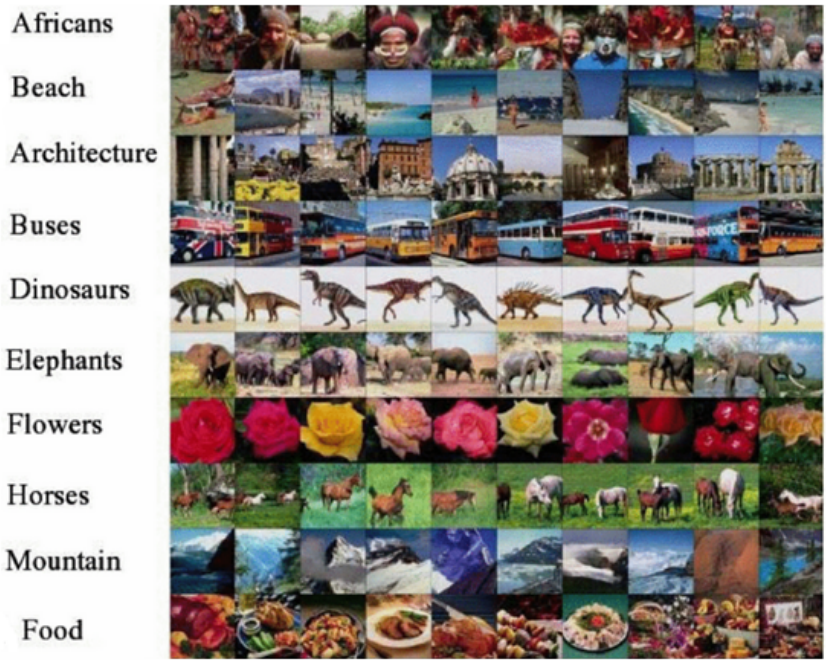

FIGURE 1.5 – Les catégories d'images présentent dans la base COREL.

COREL [6], où le problème d'annotation se limite à la détection de concepts sémantiques simples tels que montagne, éléphant comme illustré dans la figure 1.5. Néanmoins, les photos personnelles sont plus complexes et impliquent différentes personnes et divers événements (fête d'anniversaire, fête de mariage, etc). La représentation d'image par des caractéristiques de bas niveau n'est plus suffisante.

1.4 La reconnaissance d'événements dans les photos personnelles

Nous avons vu précédemment que l'annotation sémantique basée sur les caractéristiques de bas-niveau n'est plus suffisante pour décrire le contenu d'une photo vu la complexité de ce dernier. L'album photos est encore plus complexe à décrire, car il est constitué d'un ensemble de photos. Cette complexité rend le fossé sémantique très difficile à franchir. Une piste de recherche prospectée se focalise sur l'utilisation des techniques d'apprentissage pour extraire automatiquement les éléments d'une photo correspondant à des concepts visuels à partir des descripteurs de bas-niveau [3, 29]. Ainsi, à partir d'une base

6. https ://sites.google.com/site/dctresearch/Home/content-based-image-retrieval

d'apprentissage constituée d'images annotées, on peut construire des modèles capables par la suite de proposer des annotations pour de nouvelles images. Indépendamment du type d'apprentissage utilisé, l'élément à apprendre est appelé le concept et la sortie produite par un apprentissage est appelée la description du concept [30]. Cette technique a été largement utilisée dans la littérature pour la reconnaissance des objets et des scènes. Cependant, l'événement est un concept très complexe et classé à un niveau supérieur aux objets et aux scènes [31]. Les techniques d'apprentissages qui se basent directement sur les descripteurs de bas-niveau pour construire un modèle de reconnaissance d'événements ne semblent pas être la meilleure solution. Une autre piste de recherche vise à exploiter les techniques d'apprentissage automatique qui emploient des caractéristiques sémantiques pour la reconnaissance d'événements. Ces caractéristiques sémantiques représentent des descripteurs de haut niveau et correspondent aux concepts tels que les objets et les scènes.

1.4.1 Les caractéristiques sémantiques

Dans la littérature, l'annotation des images a été fortement liée à la reconnaissance de scènes et la reconnaissance d'objets. Ce problème est à la croisée des chemins de l'intelligence artificielle, la fouille de données, et la vision par ordinateur. Dans la tâche de reconnaissance visuelle, on s'intéresse à la description du contenu des images. Cette description se présente sous la forme la plus simple qui soit, c'est-à-dire une liste de mots-clés représentant les caractéristiques sémantiques et décrivant des concepts visuels présents dans les images. Ces caractéristiques ont un contenu sémantique plus riche que la simple description visuelle et peuvent, à leur tour, être employés comme des caractéristiques de haut-niveau pour prédire l'événement. La reconnaissance visuelle constitue la tâche la plus difficile en vision par ordinateur. Bien que ce domaine ait connu d'importants progrès ces dernières années, Szeliksi [32] considère que les meilleurs systèmes sont moins précis qu'un enfant de deux ans.

1.4.1.1 Les objets

La reconnaissance d'objets a suscité l'intérêt de la communauté scientifique dès les premiers pas de la vision par ordinateur. Plusieurs méthodes ont été

proposées dans ce domaine. Elles peuvent être catégorisées en trois tâches [33] :

- La catégorisation d'images qui consiste à donner un label à une image en fonction de la présence ou non d'un objet appartenant à une catégorie donnée.
- La détection d'objets qui désigne la tâche de localisation des objets d'une catégorie donnée.
- La segmentation de classes d'objets qui consiste à déterminer quels sont les pixels de l'image qui appartiennent à un objet d'une des classes d'intérêt.

Ces trois tâches sont étroitement liées. Par conséquent, les mêmes méthodes peuvent être utilisés pour les résoudre : extraction des caractéristiques visuelles, représentation locale/globale des images, construction de modèles, et appariement entre modèles et images. L'objectif de ces systèmes est de prédire la nature de l'objet dans une image au sein d'une liste exhaustive de possibilités.

1.4.1.2 Les scènes

Le but de la reconnaissance de scènes est de reconnaître le contenu d'une scène visuelle et de la classer dans une catégorie sémantique. On distingue généralement les scènes naturelles (forêt, plage, paysages de montagnes) et les scènes fabriquées par l'homme, parmi lesquelles, les scènes d'intérieur (chambre, cuisine,...) les scènes d'extérieur (ville, route,...).

La catégorisation de scènes fut une des premières approches de l'annotation automatique, avec la classification classique entre les images d'intérieur et d'extérieur, et des scènes naturelles. De nombreuses approches ont été développées afin de mieux comprendre la nature des traitements rendant possible la reconnaissance des scènes naturelles par le système visuel [34, 35, 36, 37].

Une image peut être classée dans une ou plusieurs catégories de scène. Ce type de problème est appelé multi-classification. Dans le problème multi-classification, les classes peuvent se recouvrir par définition : les images font parties intégrantes de plusieurs classes, ce qui représente la cause principale des erreurs de classification. La figure 1.6 montre un exemple de ce problème.

FIGURE 1.6 – Exemple d'images qui peuvent appartenir à deux classes différentes "croisière et Ski", problème multi-classification, tirées de la base PEC [8].

1.4.2 L'événement : un concept de haut niveau

Dans le cadre de cette thèse, nous cherchons à reconnaitre l'événement dans un album de photos personnelles. L'événement est un concept de haut-niveau qui est supérieur à l'objet et à la scène [3]. En effet, les évènements sont souvent constitués d'un ensemble d'acteurs, représentés par les objets, dans un environnement particulier, représenté par les scènes. L'évènement constitue donc l'interprétation de la présence des acteurs dans un environnement particulier. La présence du même objet dans deux environnements différents aura deux interprétations différentes. Ainsi, la présence d'une voiture dans une salle fermée entourée de gens peut être considérée comme un événement *d'exposition de voiture*, tandis que la présence d'une voiture au milieu d'une ville sera considérée comme un évènement de *visite touristique*. Si on cherche uniquement à détecter la présence ou non d'un objet en l'isolant de son environnement, l'événement sera difficile à reconnaitre. Toutefois, si on cherche à interpréter l'ensemble des acteurs et leurs environnements, la reconnaissance d'événement devient une tâche plus facile. Par conséquent, les caractéristiques sémantiques d'images jouent un rôle important dans la reconnaissance d'événements dans les albums photos. Le tableau 1.1 montre un ensemble d'événements et les caractéristiques sémantiques associées.

1.4.3 État de l'art : principaux travaux

La prolifération des appareils photos numériques a contribué à produire une grande masse de photos personnelles. Par conséquent, le développement des méthodes efficaces de gestion de photos personnelles devient crucial. Au

Événement	Objets	Scènes
Croisière	Bateau, quai	Mer, plage, côte, port
Anniversaire	Gâteau, cup cake, bougie, ballon	Restaurant, intérieure de la maison, salle des fêtes
Fête de mariage	Gâteau de mariage, robe blanche, costume, bouquet de fleurs	Salle des fêtes, jardin, lieu de culte

TABLE 1.1 – Les événements et leurs caractéristiques sémantiques.

cours des dernières décennies, de nombreux travaux de recherche ont porté sur l'organisation des collections de photos personnelles [38]. Ces travaux traitent le contenu visuel d'image pour inférer une sémantique de haut niveau telle que perçue par les humains [13]. Les chercheurs ont extrait de la sémantique à travers la reconnaissance faciale [39, 40] et l'identification des personnes [41] afin de faciliter la gestion des collections de photos. De plus, des annotations contextuelles telles que la date d'acquisition et la localisation ont également été utilisées pour le même objectif [42, 43, 44, 45].

Dans les scénarios réels, les gens prennent généralement des photos liées à des événements particuliers (par exemple, anniversaires, fête de mariage, etc.), qui sont organisées plus tard dans des albums. Selon Zigkolis [46] l'événement est considéré comme un index sémantique important pour rechercher les photos. Par conséquent, la reconnaissance automatique d'événements dans les collections de photos personnelles joue un rôle important pour la gestion intelligente des photos et la recherche avancée. Elle est également importante pour les applications telles que l'indexation sémantique d'images et la création des résumés [42, 47].

Plusieurs méthodes ont été proposées pour la reconnaissance d'événements dans les photos individuelles ou les collections de photos. Par exemple, Imran et al. [48] se sont basés sur le modèle de sac de mots pour prédire les catégories d'événements en utilisant les caractéristiques de bas niveaux. Wu et al. [49] utilisent les réseaux de neurones convolutifs pour extraire les caractéristiques

qui sont, par la suite, combinées pour prédire l'événement. Dans le travail de Roth et al. [50] une image est représentée à l'aide de caractéristiques des réseaux de neurones convolutifs (Convolutional Neural Networks -CNN [51]) et classée dans une seule catégorie d'événement culturel à l'aide du classifieur plus proches voisins itératif (Iterative Nearest Neighbors based Classifier INNC [52]). Par ailleurs, les caractéristiques contextuelles (telles que : date d'acquisition d'image, GPS, etc.,) ont également été exploitées pour améliorer la reconnaissance des événements [53, 54, 55, 56, 57].

Cependant, ces méthodes sont principalement limitées par le fait qu'elles s'appuient lourdement sur des classificateurs basés sur des caractéristiques de bas niveau. Étant donné que ces caractéristiques n'ont pas un sens sémantique explicite et peuvent être partagées par plusieurs événements, la reconnaissance des événements devient moins efficace et interprétable.

Pour résoudre ce problème, plusieurs travaux proposent d'utiliser des caractéristiques sémantiquement significatives pour la reconnaissance d'événements dans les photos. Cette solution a été inspirée initialement du domaine de la reconnaissance d'événement dans les vidéos. Beaucoup d'efforts ont été déployés pour la reconnaissance d'événements, dans les vidéos. Une grande partie de ses recherches s'orientent vers l'exploitation des caractéristiques qui ont une signification sémantique. Il s'agit d'utiliser des caractéristiques sémantiques qui décrivent le contenu de la vidéo selon une interprétation humaine. Cette technique a été appliquée dans plusieurs domaines de la vidéo [58, 59, 60]. Bien que, les travaux concernant la reconnaissance d'événements dans les vidéos basée sur les caractéristiques sémantiques sont de plus en plus nombreux, la reconnaissance d'événements en exploitant ces caractéristiques sémantiques reste un problème très peu étudié dans les photos personnelles. Cao et al. [31] ont utilisé les scènes comme caractéristique sémantique pour prédire l'événement d'une collection de photos. Ils ont exploité la corrélation entre les photos d'une collection en se basant sur la relation entre l'annotation au niveau d'événement (représenté par toute la collection, par exemple croisière) et l'annotation niveau d'image (représenté par les scènes, par exemple : mer). Quoique l'information sur les scènes puisse fournir de bons indices pour la prédiction des événements, elle reste insuffisante pour discriminer les événements qui partagent les mêmes catégories de scènes. Par exemple, les évé-

nements *fête de mariage* et *anniversaire* peuvent être associés au même type de scène *salle de fête* ; toutefois, ils peuvent contenir des objets différents. Par conséquent, les objets en avant-plan sont importants pour la reconnaissance d'événements. En d'autres termes, un événement "croisière" est généralement caractérisé par la présence de l'objet "bateau", alors qu'un événement "mariage" peut être intuitivement dérivé par la présence d'une mariée vêtue en "robe blanche" représentant l'objet d'avant-plan.

Pour résoudre ce problème, d'autres travaux ont été proposés pour la reconnaissance d'événements, en se basant sur les objets. Tsai et al. [61] ont proposé une méthode statistique qui permet de sélectionner les objets représentatifs pour des catégories d'événements prédéfinis. Un détecteur est construit pour chaque objet. L'album est représenté sur la base des réponses retournées par ces détecteurs. Cette représentation est utilisée pour prédire la catégorie d'événement en utilisant le classificateur machine à vecteur de support (SVM). Un autre travail [62], proposé par les mêmes auteurs, extrait les objets les plus fréquents pour déterminer ceux qui sont les plus discriminants. Un album photo est ensuite exprimé en fonction des fréquences de ces objets discriminatifs appelées "Compositional Object Pattern Frequency". Cependant, ces travaux ne prennent pas en considération les informations de scènes ; par conséquent, les événements contenants des objets similaires peuvent être confondus. Par exemple, les événements "tour de ville" et "expositions de voitures" peuvent tous les deux contenir l'objet *voiture* ; néanmoins, la *voiture* apparaît dans différents environnements. En d'autres termes, il est difficile de générer des classes d'événements précises en se basant uniquement sur la représentation par objets. Les informations sur l'environnement des objets sont souvent nécessaires pour comprendre le rôle des objets dans les événements [3].

Bien que plusieurs approches ont été proposées pour la reconnaissance d'événements en se basant uniquement sur la catégorisation de scènes ou la détection d'objets, la possibilité d'utiliser conjointement ces deux tâches pour reconnaître les événements n'a pas été bien étudiée. Li et al. [3] présentent un modèle graphique intégrant les informations de scènes et d'objets pour la reconnaissance d'événements dans les photos individuelles. Récemment, Wang et al. [29] proposent d'utiliser des caractéristiques dérivées de réseaux de neurones convolutifs (CNNs) pour effectuer la reconnaissance d'événements dans

des collections de photos. Sur la même lancée, Liu et al. [63] exploitent des caractéristiques de CNNs pour extraire les scènes et les objets à partir des images. Par la suite, les catégories d'événements dans les images sont prédites en utilisant une analyse discriminante. Cependant, ces méthodes sont plus adaptées à la détection d'événements dans les images individuelles. Ces dernières fournissent des informations moins riches et moins complètes sur les événements par rapport à un album entier. La variation du contenu des photos personnelles rend la tâche de description sémantique d'image plus difficile. Cette variation est due, principalement, aux raisons suivantes :

- Les caractéristiques visuelles d'une image peuvent être affectées par les conditions de sa prise, à titre d'exemple : la luminosité, l'angle et point de prise de l'image.
- La perte d'informations visuelles causée par la projection d'objets et de scène sur une image. Les objets physiques et leurs environnements sont en trois dimensions. Lorsque ils sont projetés sur une image, leur représentation est réduite en deux dimensions.
- La nature déformable des objets et des scènes.
- Les objets discriminatifs peuvent être occultés par d'autres objets (discriminatifs ou pas) dans la même scène.
- Les scènes en arrière-plan peuvent être occultées par des objets en premier-plan.
- Dans un album, plusieurs photos peuvent ne pas contenir des scènes ou de photos clés qui portent une information utile sur l'événement.

La variation visuelle crée le problème de l'ambiguïté visuelle d'une photo. Chaque variation mène à une apparence visuelle différente. Par ailleurs, certaines photos dans un album peuvent ne pas contenir des objets et scènes discriminatifs ce qui signifie qu'elle n'apporte aucune information pertinente sur l'événement. La considération d'un ensemble de photos pour la prédiction d'un événement peut atténuer l'acuité de ces problèmes. D'une part, les photos constituant un album sont généralement prises à des moments reflétant des moments importants des événements [38] et, par conséquent, peuvent

être plus informatives. Par ailleurs, les photos constituant un album peuvent donner une vue exhaustive sur l'ensemble des objets/scènes impliqués dans l'événement.

1.4.4 Discussion et positionnement de l'approche proposée

Nous présentons dans cette section une étude comparative entre les différentes approches existantes pour la reconnaissance d'événements dans les albums. Nous résumons cette dernière à travers un tableau comparatif entre ces approches. Enfin, nous proposons une discussion et nous positionnons notre approche par rapport à l'existant. Le tableau comparatif 1.2 résume comment les différentes approches ont procédé à la résolution du problème de la reconnaissance d'événements dans les photos. Quatre critères sont utilisés pour la comparaison : le type de caractéristiques utilisées pour la représentation des images (descipteurs de bas niveau, objet, scène, ou contextuelle), le nombre d'images considérées (image individuelle/ensemble d'images), la combinaison des objets et de scènes, et l'utilisation de la notion de pertinence des caractéristiques.

Pour pallier les limitations des solutions existantes, nous proposons une approche qui combine les objets et les scènes pour la reconnaissance d'événements dans les albums photos. Plus précisément, un réseau bayésien est proposé pour inférer la catégorie d'événements, en tirant profit des informations sur les objets et les scènes extraites à partir d'un ensemble d'images. Étant donné un album d'images, nous utilisons d'abord les réseaux de neurones convolutifs pour extraire les objets et les scènes de chaque image de l'album. Les objets et scènes extraits constituent les caractéristiques sémantiques de niveau moyen pour la reconnaissance d'événement. Le réseau bayésien proposé modélise les relations entre les catégories d'événements, les objets et les scènes. Dans le même modèle, nous proposons d'intégrer la pertinence des caractéristiques objet/scène pour améliorer la reconnaissance d'événements. Enfin, pour déduire la catégorie d'événement d'un nouvel album, nous combinons les caractéristiques de ses images par une fonction de vraisemblance et nous utilisons la probabilité a posteriori maximale (MAP) pour estimer les catégories d'événements.

Contrairement aux approches existantes pour la reconnaissance d'événe-

Approche reconnaissance d'événement	Type de caractéristiques utilisées	Image individuelle	Ensemble d'image	Combinaison d'objet et de scène	Pertinence de caractéristiques
Naaman et al. [44]	Contextuelles	-	+	-	-
Kisilevich et al. [43]	Contextuelles	-	+	-	-
Zigkolis et al. [46]	Contextuelles	-	+	-	-
Dao et al. [53]	Contextuelles et visuelles	-	+	-	-
Li et al. [3]	Objet et scène	+	-	+	-
Wang et al. [29]	Objet et scène	+	-	+	-
Imran et al. [48]	Visuelles	-	+	-	-
Wu et al. [49]	Visuelles	-	+	-	-
Cao et al. [31]	Scène	-	+	-	-
Tsai et al. [61]	Object	-	+	-	-
Tsai et Cao [62]	Object	-	+	-	-
Bossard et al. [38]	Contextuelles et scène	-	+	-	-
Rothe et al. [50]	Visuelles	+	-	-	-
Notre Approche	Objet et scène	+	+	+	+

TABLE 1.2 – Étude comparative entre les approches d'annotation de photos personnelles pour la reconnaissance d'événements.

ments dans les albums [38, 53], notre approche repose uniquement sur les informations visuelles. Le temps ou la localisation peuvent être intégrés dans des versions ultérieures de ce travail pour améliorer les performances. Nous prenons en considération toutes les photos d'un album pour avoir une vue complète sur l'événement. De plus, nous combinons les objets et les scènes et nous introduisons leurs pouvoirs discriminatifs (c'est-à-dire pertinence) pour une reconnaissance d'événements plus précise. Par exemple, l'objet "citrouille" peut aider à déterminer l'événement "Halloween", mais pas l'événement "Ski". De même, l'objet "arbre de Noël" apporte beaucoup d'information sur la catégorisation de l'événement "Noël" que les autres objets. Par conséquent, l'objet 'arbre de Noël' est considéré comme une information pertinente pour prendre la bonne décision sur la catégorisation d'événement "Noël". La pertinence objet/scène peut être un outil très efficace pour avoir une reconnaissance d'événements plus précise. Nos principales contributions peuvent être résumées comme suit :

- Un album photo est constitué d'un ensemble de photos qui représente

un événement particulier. Notre approche permet de prédire la classe d'événement pour un album photos. Nous avons vu que l'analyse d'un ensemble de photos apporte des informations plus riches sur un événement par rapport à l'analyse d'une image individuelle. Comme dans une histoire ou un StoryTelling, l'événement d'un album est décrit par l'ensemble des images où chaque image représente un épisode.

- Nous utilisons les caractéristiques sémantiques pour représenter un album. Les caractéristiques sémantiques de niveau-moyen, constituées de scènes et d'objets, aident à combler le fossé sémantique et avoir une représentation plus interprétable au niveau de l'utilisateur.

- Nous introduisons un schéma de pertinence de caractéristiques qui incorpore la puissance prédictive de chaque objet/scène pour la reconnaissance d'événement. L'utilisation de la pertinence est motivée par plusieurs découvertes d'études psychologiques dans la cognition humaine, où la compréhension de la scène dans le monde réel repose en grande partie sur l'analyse des objets et leurs contextes [64, 65, 66, 67]. Le plus souvent, les gens peuvent déduire directement leur compréhension sur les événements après avoir vu certains objets se produire dans des contextes spécifiques [68, 69]. Cette logique est mise en évidence en considérant la pertinence des caractéristiques pour stimuler la reconnaissance humaine d'événements.

- Nous proposons un réseau bayésien pour la reconnaissance d'événements dans les albums photos en combinant les scènes et les objets. La prédiction d'événements pour les nouveaux albums est faite en utilisant l'inférence bayésienne. Depuis les travaux de Rand [4], plusieurs études ont discuté comment la perception de l'événement se produit dans le système de vision humaine (en anglais Humain Vision System, HVS) [70, 71, 72]. L'avènement des méthodes bayésiennes a offert un nouvel outil attrayant pour explorer le potentiel de la modélisation de l'inférence visuelle de manière plus proche du HVS. Par conséquent, les réseaux bayésiens représentent un outil prometteur à explorer pour faciliter la reconnaissance des événements dans les ensembles d'images. Par ailleurs, l'utilisation des modèles graphiques probabiliste dans le domaine de re-

connaissance d'évènements sportifs dans les images a donné des résultats motivants[3] qui nous ont encouragé à explorer l'application des réseaux bayésiens pour la reconnaissance d'évènements dans les albums photos.

1.5 Conclusion

Dans ce chapitre, nous avons présenté les concepts de base liés à la recherche et l'annotation sémantique d'images. Nous nous sommes focalisés sur l'annotation automatique d'images avec ses deux variations : les approches basées sur les caractéristiques contextuelles et les approches basées sur les caractéristiques visuelles. Puis, nous avons souligné les limites des systèmes d'annotations basées sur les descripteurs contextuels et visuels dans le domaine de la reconnaissance d'événements dans les images. Nous avons également présenté les caractéristiques sémantiques d'une image en nous nous focalisons sur la reconnaissance d'objet, la reconnaissance de scène et leurs rôles dans la reconnaissance d'événements. Enfin, nous avons présenté l'évolution du domaine et nous avons positionné notre travail par rapport à l'existant.
Dans le prochain chapitre, nous exposons quelques techniques d'apprentissage automatique qui sont nécessaires à la compréhension de l'étude menée.

Chapitre 2
APPRENTISSAGE AUTOMATIQUE

2.1 Introduction

Le développement actuel des ordinateurs à l'échelle mondiale crée de nouveaux besoins et permet à des applications innovantes d'apparaître. Ainsi, la croissance exponentielle de nombre d'images numériques donne lieu à une multitude de nouvelles applications, comme la reconnaissance de visage, l'aide au diagnostic médical, la catégorisation automatiques des photos, etc. Ces problèmes peuvent être aisément résolus par un être humain mais ils sont très couteux en terme de temps. Il est aujourd'hui possible de résoudre ces problèmes de manière automatique grâce au *machine learning* appelé aussi *apprentissage automatique*. Les techniques d'apprentissage automatique sont nombreuses. Cependant, il n'existe pas un algorithme qui convient à tous les types de problèmes. Par ailleurs, il n'existe pas d'algorithmes universellement plus performants et efficaces que d'autres.

Dans ce chapitre, nous allons présenter les types d'apprentissage, qui diffèrent l'un de l'autre selon leurs objectifs. Nous allons voir l'apprentissage supervisé et l'apprentissage non-supervisé. Nous mettons l'accent sur l'apprentissage supervisé avec ses deux types : discriminatif et génératif et nous détaillons plus précisément les approches discriminatives, et nous discutons les approches génératives dans le chapitre prochain. Nous allons voir comment améliorer les performances des algorithmes d'apprentissage par sélection de caractéristiques.

2.2 Apprentissage automatique

A l'intersection entre l'intelligence artificielle et les statistiques, l'apprentissage automatique a pour but la résolution automatique des problèmes à partir d'exemples passés. Vincent [73] définit l'apprentissage automatique par "une tentative de comprendre et de reproduire l'habileté humaine d'apprendre de ses expériences passées et de s'adapter dans les systèmes artificiels". L'ap-

prentissage est la capacité de généraliser et de résoudre de nouveaux cas à partir des exemples réussis dans le passé. L'apprentissage automatique est une branche connexe à l'intelligence artificielle puisqu'il puisait initialement ses sources des neurosciences. Il fait appel aussi à des théories et outils d'autres disciplines tels que : théorie de l'information, traitement du signal, statistique [74].

L'apprentissage automatique est appliqué dans les cas où il est difficile ou impossible de remplir des tâches par des moyens algorithmiques classiques. En d'autres termes, l'apprentissage automatique permet de concevoir des modèles prédictifs à partir d'un corpus de données. Plusieurs applications tirent profit de l'apprentissage automatique. Parmi ces applications, nous pouvons citer : le filtrage de pourriel [75], la reconnaissance de visages [76], la reconnaissance des empreintes digitales [77] et la reconnaissance de la parole [78]. Par exemple, dans le cas de la reconnaissance de la parole, les algorithmes d'apprentissage permettent de construire un modèle à partir d'un ensemble de sons et leurs retranscriptions en langue naturelle dit "ensemble d'apprentissage". Le modèle construit est capable de donner sa propre transcription d'une nouvelle séquence sonore qui n'appartient pas à l'ensemble initial des sons.

Plusieurs problèmes sont rencontrés dans l'apprentissage automatique, parmi lesquels le sur-apprentissage (Overfitting). Ce problème est défini comme une erreur de modélisation qui se produit lorsque le modèle apprend une fonction trop conforme à un ensemble de données limité [79]. En réalité, les données étudiées ont souvent un certain degré d'erreur ou de bruit aléatoire. Par conséquent, le modèle appris sera difficilement généralisé face à de nouvelles données. La généralisation désigne la capacité d'un modèle à pouvoir effectuer des prédictions robustes sur des nouvelles données. De plus, la tentative de rendre le modèle conforme de trop près à un ensemble limité de données inexactes peut infecter le modèle avec des erreurs substantielles et réduire son pouvoir prédictif. Par ailleurs, pour une meilleure capacité de généralisation, on s'efforcera toujours de faire en sorte que le nombre de paramètres ajustables d'un modèle soit petit devant le nombre d'exemples.

Afin d'évaluer l'efficacité de l'algorithme d'apprentissage, la validation croisée (Cross Validation) est largement utilisée pour évaluer les performances

d'un modèle à prédire de nouvelles données. Cette technique consiste à diviser l'ensemble des données d'apprentissage en deux sous-ensembles : le premier sous-ensemble est utilisé pour construire le modèle tandis que le deuxième sous-ensemble est employé pour l'évaluer. Il existe plusieurs techniques de validation croisée [80] :

- **Testset validation**, elle consiste à diviser l'ensemble d'apprentissage de taille n en deux sous ensembles, Le premier ensemble d'apprentissage (communément supérieur à 60%) et le deuxième ensemble pour les tests. Le modèle est appris sur l'ensemble d'apprentissage et validé sur le deuxième ensemble de test. L'erreur de prédiction ou les mesures de performances est estimée sur l'ensemble de test selon le type de problème (e.g. la précision, l'erreur quadratique moyenne)

- **k-fold cross validation**, elle consiste à diviser l'ensemble d'apprentissage en k ensembles. Par la suite, un des k ensembles est sélectionné comme ensemble de validation et les $(k-1)$ ensembles restants sont utilisés pour apprendre le modèle. L'opération est répétée k fois, en sélectionnant à chaque itération un ensemble de validation non pris en compte dans les itérations précédentes. Les mesures de performances/erreurs calculées sont moyennées au fur des itérations pour estimer l'erreur ou la mesure de performance finale.

- **Leave-one-out cross validation**, elle est un cas particulier de la deuxième méthode où $k = n$ [81], c'est-à-dire que l'on apprend sur $(n-1)$ observations puis on valide le modèle sur la énième observation et l'on répète cette opération n fois.

La validation croisée est souvent utilisée pour éviter le problème du sur-apprentissage [82]. Elle permet aussi l'optimisation des paramètres des méthodes d'apprentissage [83].

2.3 Types d'apprentissage automatique

Il existe plusieurs types d'apprentissages automatique ; ils se distinguent principalement par la nature des données qui doivent être apprises. On focalise

notre étude sur l'apprentissage supervisé et l'apprentissage non-supervisé [84].

- Dans le cas de **l'apprentissage supervisé** [85], l'objectif de l'apprentissage est déterminé explicitement via la définition d'une cible à prédire. Dans ce cas, $\mathcal{D}$ correspond à un ensemble de données d'entrainement constitué de n paires d'entrée $X^{(i)}$ et de cibles $Y^{(i)}$, tel que : $\mathcal{D} = \{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots, (X^{(n)}, Y^{(n)})\}$. Autrement dit, on dispose d'un ensemble d'apprentissage annoté et connu. Il existe plusieurs algorithmes d'apprentissage supervisé, parmi lesquels nous citons les arbres de décisions, les réseaux de neurones, les machines à vecteurs de supports (SVM), etc.

- Dans le cas de **l'apprentissage non-supervisé** [85], on dispose d'un nombre fini de données d'apprentissage sans aucune étiquette. L'ensemble d'apprentissage $\mathcal{D}$ contient seulement les descriptions des données d'apprentissage et aucune cible n'est prédéterminée, tel que : $\mathcal{D} = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$. L'apprentissage non-supervisé est le plus souvent utilisée pour des applications de regroupement "clustering", dont le but est de construire des groupes homogènes, représentants des classes de données, à partir de l'ensemble d'apprentissage $\mathcal{D}$. Dans ce processus, il revient aux algorithmes de découvrir par eux-mêmes les structures plus ou moins cachées des données, afin de former des groupes d'individus dont les caractéristiques sont communes. Les méthodes de classification non supervisées peuvent être réparties en deux grands groupes à savoir : les méthodes de partitionnement et les méthodes hiérarchiques. En général, toutes ces méthodes s'appuient sur la notion de similarité intraclasse et de dissimilarité interclasses.

Ces différents types d'apprentissage peuvent être aussi combinés dans un seul système. Cette combinaison donne naissance à une autre approche d'apprentissage automatique dite approche *semi-supervisée* [86]. Dans cette dernière, les données de l'ensemble d'apprentissage $\mathcal{D}$ ne sont pas toutes annotées, souvent un petit nombre de données annotées et une grande quantité de données non étiquetées, tel que : $\mathcal{D} = \{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots, (X^{(m)}, Y^{(m)}), X^{(m+1)}, X^{(2)}, \ldots, X^{(n)}\}$ où $1 < m < n$. Cet apprentissage est considéré comme un type intermédiaire entre

l'apprentissage supervisé et l'apprentissage non-supervisé. Nous sommes dans un cadre d'apprentissage supervisé si les sorties sont observées, non supervisé si elles ne le sont pas et semi-supervisé si elles ne sont que partiellement observées.

Dans le domaine d'annotation automatique d'images, la sortie correspond à des mots clés représentant des concepts visuels et décrivant l'image. Ainsi, un modèle est construit à partir d'une base d'apprentissage d'images annotées. On se situe donc dans la branche de l'apprentissage supervisé qui a pour but de classer des nouvelles observations dans une catégories prédéfinie. On peut distinguer deux types d'approches pour résoudre un problème d'apprentissage supervisé [87] : *l'approche discriminative* et *l'approche générative*. Dans ce qui suit nous présentons l'approche discriminative, tandis que l'approche générative sera détaillée dans le chapitre suivant.

2.4 Approches discriminatives

Dans l'apprentissage supervisé, les approches discriminatives utilisent les données d'apprentissage pour construire directement une correspondance entre les entrées X et les classes cibles Y [88]. De manière générale, on désigne par approches "discriminatives" les méthodes qui ne s'intéressent qu'aux relations d'entrée-sortie lors de l'apprentissage. Le k-plus proches voisins et les machines à vecteurs de supports sont parmi les méthodes discriminatives les plus utilisées pour l'apprentissage supervisé [89].

2.4.1 Le k plus proches voisins

Le k plus proches voisins, souvent abrégé en k-*ppv* (k-plus proches voisins) [90], consiste à rechercher parmi l'ensemble d'apprentissage, un nombre k d'exemples parmi les plus proches possibles de l'exemple à classer. Par la suite, le nouvel exemple est affecté à la classe majoritaire parmi ces k exemples trouvés. Par exemple, si $k = 1$, alors le nouvel exemple est affecté à la classe du plus proche voisin de l'ensemble d'apprentissage. Le nombre k est fixé a priori par l'utilisateur [91]. La figure 2.1 représente un exemple de classification avec la méthode $k - ppv$ avec $k = 4$. Dans cet exemple, les quatre plus proches voisins de a sont b^4, b^2, b^5, b^3. L'exemple a sera affecté à la classe majoritaire parmi ces quatre voisins. Il convient de souligner qu'il est possible d'utiliser

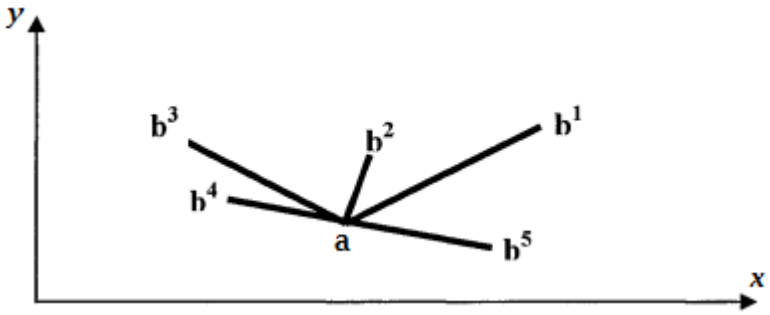

FIGURE 2.1 – Méthode 4-ppv.

un seuil au-dessous duquel une décision de rejet est prise. La méthode *k-ppv* est une méthode simple à mettre en œuvre. Par ailleurs, elle n'impose aucune hypothèse a priori sur les données. Cependant, la qualité de la discrimination dépend du choix du nombre k de voisins considérés. Cette méthode présente plusieurs inconvénients [92]. Elle nécessite une mémorisation de l'ensemble de données d'apprentissage ce qui la rend gourmande en mémoire. De plus, elle calcule les distances entre l'exemple de test et l'ensemble des exemples d'apprentissage, ce qui augmente le temps de calcul.

2.4.2 Les machines à vecteurs de supports

Les Machines à vecteurs de supports (SVMs) est une méthode développée par Vapnik [93, 21]. Il reste l'un des algorithmes les plus utilisés par la communauté de vision par ordinateur dû a sa capacité de généralisation. L'algorithme SVM a été conçu, au départ, pour la classification binaire [94]. Le principe de SVM consiste à trouver un hyperplan séparateur distinguant parfaitement les deux classes de données, en assurant le principe de maximisation de la marge séparant les données de cet hyperplan. La marge est la distance qui sépare la frontière de séparation ($f(x) = 0$) des exemples les plus proches ; ces derniers sont appelés *vecteurs de supports*. Les données situées à l'extérieur de ces marges ne participent pas à la définition de la frontières de décision. La figure 2.2 illustre le principe général des SVMs dans le cas binaire.

Généralement, les données ne sont pas linéairement séparables. Les SVMs s'appuient sur la notion de noyau qui les rend mieux adaptés pour résoudre les problèmes des données non linéairement séparables [95]. Cette notion permet

toutefois de projeter les données dans un espace de dimension supérieure, dans lequel les données pourront être séparables linéairement. Il existe plusieurs types de noyau [96] :

- Noyau linéaire [96] : Si les données sont linéairement séparables, on n'a pas besoin de changer d'espace, et le produit scalaire suffit pour définir la fonction de décision : $K(x_i, x_j) = x_i^T x_j$
- Noyau polynomial [96] : Le noyau polynomial élève le produit scalaire à une puissance naturelle d : $K(x_i, x_j) = (x_i^T x_j)^d$ Si $d = 1$ le noyau devient linéaire.
- Noyau Sigmoïde [97] : $K(x_i, x_j) = tanh((x_i^T x_j) + c)$

Avec x_i, x_j deux vecteurs différents.

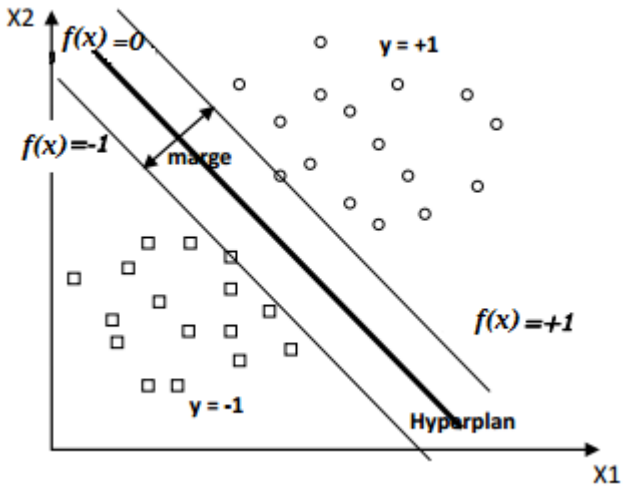

FIGURE 2.2 – SVM binaire [96].

Bien que les machines à vecteur support sont à leurs origines binaires, elles ont été étendues pour des cas multiclasses [94]. Dans le monde réel, la plupart des problèmes sont de type multiclasses, l'un des exemples les plus connus est la reconnaissance de scènes. Dans ce type de problème, un nouvel exemple est associé à une classe parmi plusieurs au lieu d'être associé à une des deux classes comme dans le cas binaire. Par conséquent, un seul

hyperplan n'est plus suffisant car la décision n'est plus binaire. Afin de pallier à ce problème, les machines à vecteurs de supports réduisent le problème multiclasses à une composition de plusieurs hyperplans biclasses permettant de tracer les frontières de décision entre les différentes classes [98]. Dans ces méthodes, l'ensemble de données est décomposé en plusieurs sous ensembles représentant chacun un problème de classification binaire. Un hyperplan de séparation est créé pour chaque problème binaire. On trouve dans la littérature, deux approches qui présentent des solutions palliatives aux problèmes de classification multiclasses par la décomposition de l'ensemble des exemples [99] :

- **Une-contre-reste (one-versus-all)** : Cette méthode vise à construire pour chaque classe un classifieur binaire pour la séparer de toutes les autres classes [99]. Pour un problème de k classes, k classifieurs SVMs binaires seront construits comme illustré en figure 2.3. Pour affecter un exemple x à l'une des k classes, la méthode "*une-contre-reste*" utilise le principe de "le gagnant prend tout" (winner-takes-all) pour prendre une décision, où la classe retenue est celle qui est associée au classifieur ayant retourné le score le plus élevé.

- **Une-contre-une (one-versus-one)** : Cette méthode, comme illustrée en figure 2.4, vise à apprendre un classifieur par paires de classes [99]. Pour k classes, les $k(k-1)/2$ classifieurs seront construits, en regroupant les classes deux à deux. Pour affecter un exemple x à une classe, la méthode "*une-contre-une*" effectue un "vote majoritaire" (max wins voting), où chaque classifieur renvoie un vote associé à la classe qu'il a reconnu. Une fonction du score peut être utilisée pour pondérer le vote de chaque classifieur. Le nouvel exemple x est affecté à la classe la plus votée.

2.4.3 Les réseaux de neurones

2.4.3.1 Perceptrons multi-couches

Un perceptron multi-couches (PMC, en anglais Multi-Layer Perceptron, MLP), appelé aussi réseau de neurones multi-couches, peut être vu comme un ensemble d'unités de traitement, appelés nœuds ou neurones, reliées entre

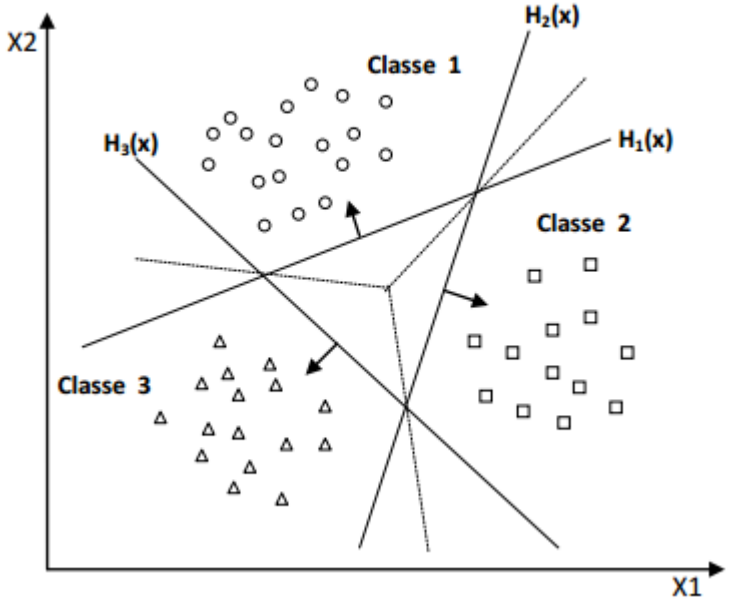

FIGURE 2.3 – Méthode une-contre-reste [96].

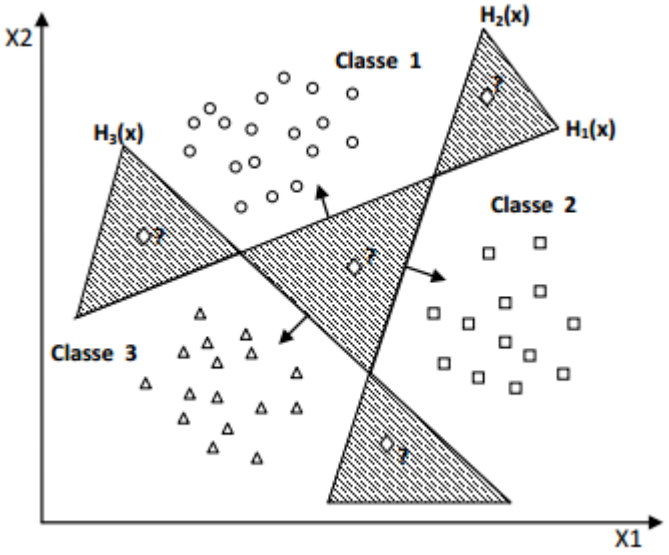

FIGURE 2.4 – Méthode une-contre-une [96].

elles par des connections pondérées [100]. Les poids de ces connections sont les paramètres du modèle. Ces neurones et ces connections sont organisés

en couches. La première couche est appelée couche d'entrée, la dernière est appelée couche de sortie et la ou les couches du milieu sont appelées couches cachées comme l'illustre la figure 2.5.

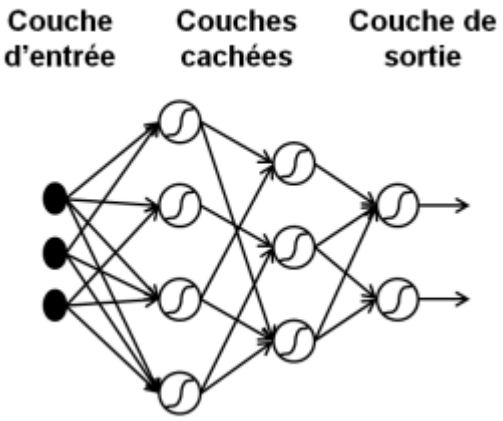

FIGURE 2.5 – Exemple d'un perceptron multi-couches avec une couche d'entrée, deux couches cachées et une couche de sortie [100].

Les neurones de ces couches cachées, ainsi que ceux de la couche de sortie appliquent deux traitements : une combinaison linaire de leurs entrées (dont les poids sont des paramètres du réseau), suivie d'une fonction non-linéaire appelée fonction d'activation [101]. Les deux fonctions les plus courantes sont la *tangente hyperbolique* et la *fonction sigmoïde*. Dans ce qui suit $\sigma_i(x)$ désignera la fonction d'activation correspondant à un neurone i. Si nous considérons un perceptron multi-couches avec N neurones d'entrée, activés par un vecteur d'entrée x (de taille N), et par $w_{i,j}^{0,1}$ le poids correspondant à la connexion entre le neurone i de la couche 0 et le neurone j de la couche 1, la sortie a_j^1 de chacun des neurones de la première couche cachée sera exprimée par [100] :

$$a_j^1 = \sigma_j \left(b_j^1 + \sum_{i=1}^{N} w_{i,j}^{0,1} x_i \right) \tag{2.1}$$

où σ_j est la fonction d'activation décrite précédemment, et b_j^1 est un paramètre supplémentaire appelé biais, qui peut être considéré comme le poids d'une entrée constante égale à 1, et dont le rôle est de rajouter un degré de liberté supplémentaire en agissant sur la position de la frontière de décision

(pour plus de détails, se référer à [102]). Ce même processus exprimé par 2.1 peut être répété pour les autres couches (cachées ou celle de sortie) : chaque sortie d'une couche l joue le rôle d'entrée pour la couche suivante $l+1$. Les perceptrons multi-couches sont généralement utilisés pour des problématiques de classification supervisées. Ceci implique l'existence d'un ensemble de paires d'entrées-sorties (appelé base d'apprentissage) liés par une certaine relation, que le réseau va apprendre en ajustant ses paramètres.

2.4.3.2 Réseaux de neurone à convolution

Les réseaux de neurones à convolution, appelés aussi ConvNets, (Convolutional Neural Networks-CNN) ont suscité un grand intérêt parmi les communautés de vision par ordinateur et de classification d'images. Récemment, les CNNs ont été largement étudiés et appliqués pour la classification des images, la reconnaissance des objets [103], la reconnaissance des scènes [104], etc. Les architectures des réseaux de neurones à convolution sont constituées de plusieurs couches de modules non-linéaires paramétrisés, et qui sont soumises à l'apprentissage. Chaque couche, dans l'architecture profonde, permet une représentation de plus haut niveau que celle qui la précède.

Un réseau de neurones convolutif contient deux couches en alternance : une couche de convolution et une couche de sous-échantillonnage (pooling) [79]. La couche de convolution est composée d'un certain nombre de filtres (appelées aussi noyaux). Ces filtres permettent de détecter des caractéristiques élémentaires sur l'image. Ils sont définis par des poids qui sont ajustables durant l'apprentissage. La couche de convolution opère sur son entrée une convolution dont les filtres sont les poids à apprendre, et alimente la couche suivante. La convolution des filtres peut être vue comme un partage de poids qui représente l'idée clé des réseaux de neurones à convolution. Le principe de partage de poids consiste à appliquer les unités d'un même filtre à différents positions de l'image, ce qui permet de réduire la complexité en diminuant le nombre de paramètres à apprendre.

La figure 2.6 présente la structure d'une couche de convolution. Les sorties de la couche de convolution sont transmises à la couche de sous-échantillonnage [79]. Cette dernière réduit la dimension de la couche précédente grâce à une fonction de sous-échantillonnage (généralement la moyenne

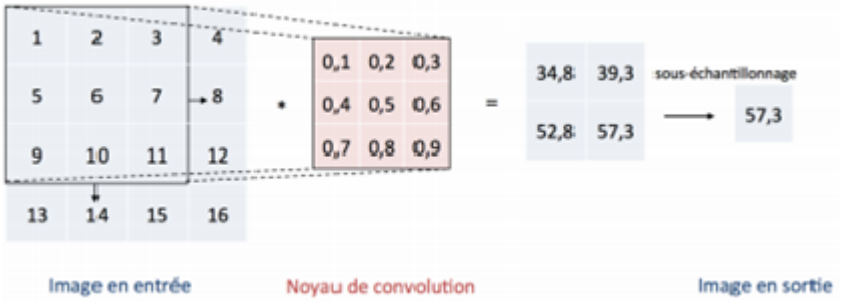

FIGURE 2.6 – Principe d'une couche de convolution dans un réseau de neurones convolutif [100]. Dans cette figure, l'image en entrée (couche d'entrée) a une largeur et une longueur. Chaque filtre (ou noyau de convolution) de taille 3×3 est appliqué à toute l'image de taille 4×4 avec des poids similaires par la technique de fenêtre glissante. L'image produite est de taille 2×2. Un sous-échantillonnage de taille 2×2 (où la valeur maximale est conservée) génère une image de sortie de taille 1×1.

ou le maximum des sorties de la couche de convolution précédente). La couche suivante effectue ensuite une convolution sur les sorties de la couche de sous-échantillonnage, et ainsi de suite. À la dernière couche de sous-échantillonnage, une couche de neurones "classiques" est rajoutée à la sortie pour permettre d'effectuer l'entraînement supervisé et la classification des données encodées. L'alternance entre les couches de convolution et les couches de sous-échantillonnage donne une structure pyramidale au CNN. Le principal objectif de l'apprentissage profond est de construire automatiquement de l'image en entrée, une représentation de plus en plus de haut-niveau de couche en couche.

La figure 2.7 illustre un exemple de structure d'un réseau de neurone à convolution LeNet [105]. Les cartes de caractéristiques sont représentés en gris et notés par des C_i, les cartes de sous-échantillonnage sont représentés en bleu et notés par des S_i, et les neurones sont représentés par des ronds blancs et notés par des N_i.Toutefois, le nombre de couches, celui des cartes (filtres) ainsi que leurs dimensions sont des paramètres architecturaux qui varient d'une problématique à une autre.

Le réseau LeNet [105] est constitué de sept couches cachées, en plus d'une couche d'entrée. Ces couches cachées peuvent être classées en deux catégo-

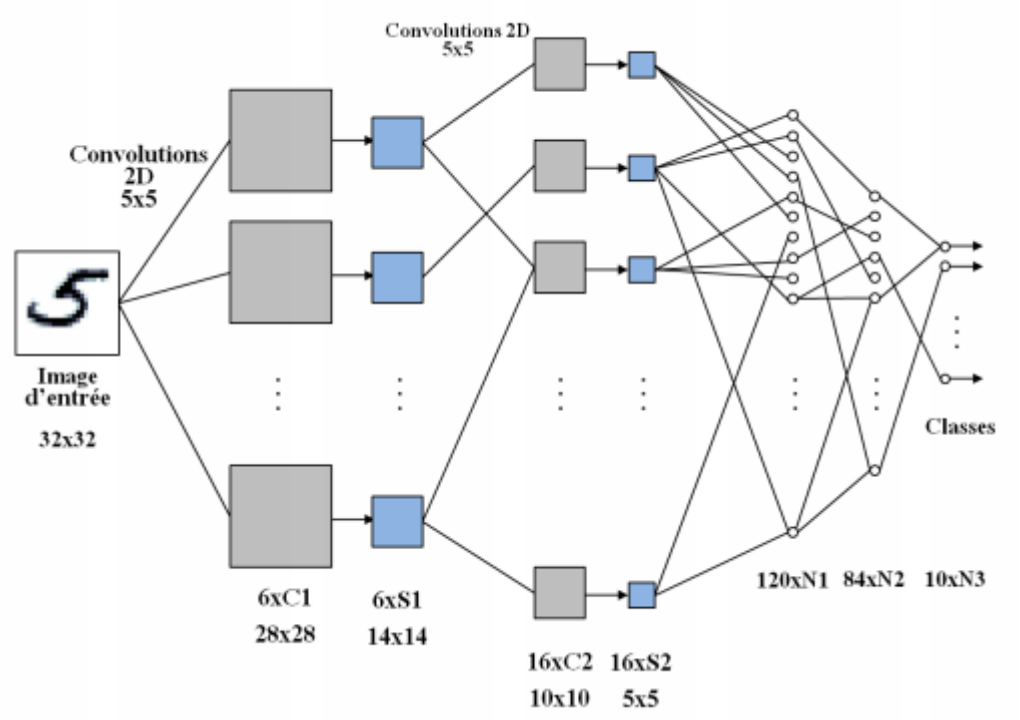

FIGURE 2.7 – Architecture du réseau de neurones à convolutions LeNet [105].

ries : les quatre premières couches sont des cartes de caractéristiques (ou de sous-échantillonnage),et les trois dernières couches sont des neurones classiques (similaires à celles d'un perceptron multicouche), où chaque neurone est connecté à tous les neurones de la couche précédente.

Des architectures de réseaux de neurones convolutifs plus complexes, comme *AlexNet* [106], *ZF-Net* [107], *VGG* [108] , *GoogleNet* [109] et ResNet [1], ont été proposées dans la littérature. Ces réseaux de neurones convolutionnels ont été comparés sur la base ImageNet et diagramme montre les résultats obtenus par ces réseaux. Selon le diagramme 2.8, les architectures tendent à devenir de plus en plus profondes au fur des années. Malgré que le réseau ResNet a obtenu les meilleurs résultats sur la base ImageNet avec un taux d'erreur de 3.57% par rapport à l'erreur commise par l'être humain de 5.1%, il reste une architecture très profonde à adopter et qui nécessite un matériel très puissant pour son entrainement.

Le réseau *GoogleNet*, proposé par Szegedy et al., a obtenu les meilleurs

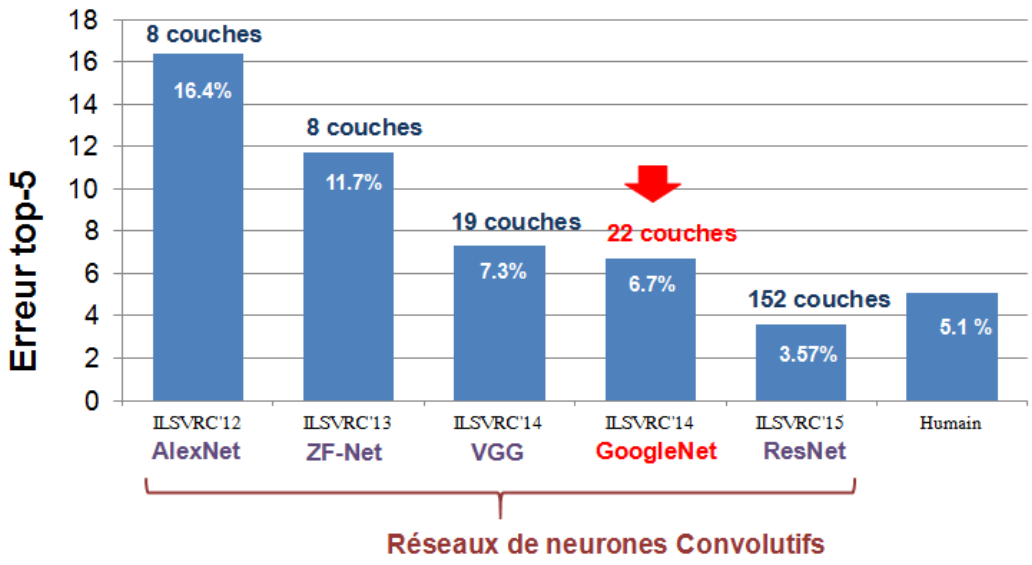

FIGURE 2.8 – Résultats obtenus sur la base du challenge ImageNet LSVRC(ImageNet Large Scale Visual Recognition Challenge [1].

résultats lors du défi ILSVRC 2014 pour la reconnaissance d'objets avec un taux d'erreur de 6.7%. Le réseau est constitué de 22 couches qui reste facile à adopter. Sa principale contribution a été le développement d'un module qui a considérablement réduit le nombre de paramètres dans le réseau (4 millions de paramètres pour GoogleNet par rapport à *AlexNet* avec 60 millions), nommé Inception. Les blocs Inception sont une combinaison de convolution multiple et de sous-échantillonnage exécutées sur les mêmes volumes d'entrée. Les sorties de toutes ces opérations sont concaténées pour former un seul volume de sortie, ce qui élimine une grande quantité de paramètres qui ne semblent pas beaucoup importants. La figure 2.9 montre une structure d'un bloc Inception.

Cependant, la création d'un modèle de réseau de neurones convolutif est coûteuse en terme de matériel et de quantité de données d'apprentissage nécessaires [110]. De plus, trouver la bonne architecture du réseau de neurones convolutif est une tâche complexe. Il faut d'abord fixer le nombre de couche et la tailles et le nombre de filtres qui seront utilisés dans les opérations matricielles. L'entraînement consiste alors à en optimiser les coefficients pour minimiser l'erreur de classification. Cet entraînement peut prendre plusieurs semaines pour une architecture CNN, avec de nombreux GPUs travaillant sur

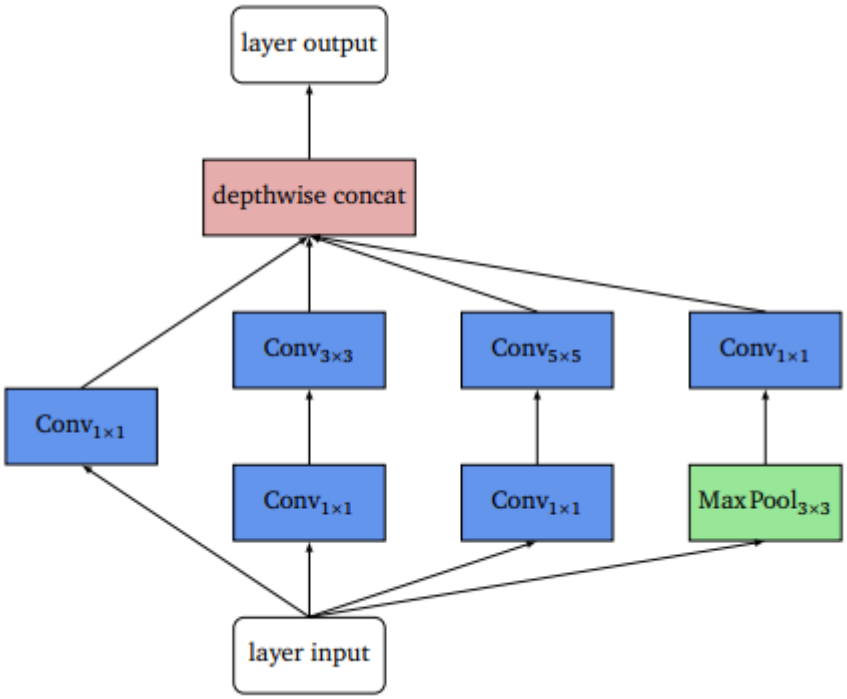

FIGURE 2.9 – bloc Inception. L'entrée est traitée de manière indépendante par quatre différentes pipelines, qui créent tous les mêmes dimensions spatiales. La sortie du bloc Inception est créée à partir de la concaténation des résultats des pipelines [109].

de millier d'images. Toute la complexité de création d'un réseau CNN peut être évitée en adaptant un réseau pré-entraîné disponible. Une telle technique est appelées *Transfert learning* [110], car on exploite la connaissance acquise sur un problème de classification général pour l'adopter à un problème de classification particulier. Dans ce qui suit, nous allons nous intéresser à la technique de *Transfer learning*.

2.4.4 Transfer learning

La connaissance sur la classification d'images contenue dans un réseau convolutif pré-entraîné peut être exploitée de deux façons [111], comme l'illustre la figure 2.10 : **extracteur automatique de caractéristiques d'images** et **Fine Tuning**. Nous commençons par détailler comment utiliser un réseau convolutif pour extraire les caractéristiques de bas niveau. Par la suite, nous présentons la méthode Fine tuning.

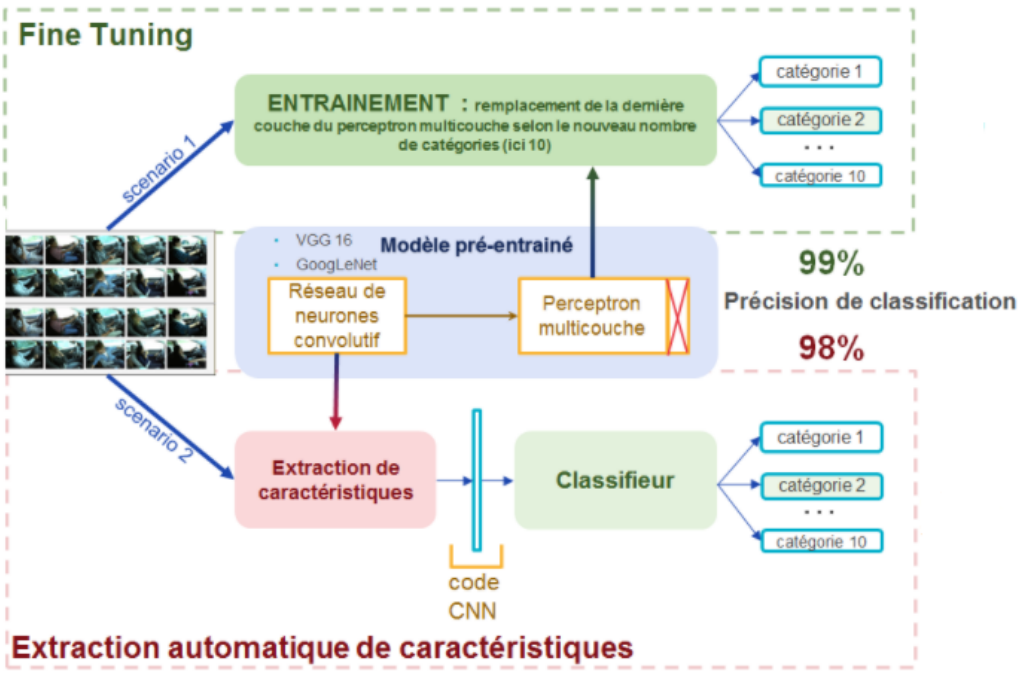

FIGURE 2.10 – Réutilisation d'un réseau de neurones pré-entraîné.

2.4.4.1 Extracteur automatique de caractéristiques d'images

A partir de l'architecture d'un réseau convolutif, des caractéristiques de bas niveau peuvent être extraites pour représenter les images. Pour l'extraction automatique des caractéristiques, seulement la partie convolutive d'un réseau pré-entraîné est exploitée. Les caractéristiques extraites seront utilisées pour alimenter un classifieur, comme illustré à la figure 2.10. Toutefois, les caractéristiques extraites sont de bas niveau et ne peuvent être interprétées au niveau sémantique. Dans une architecture de réseau convolutionnel, la première partie d'un CNN fonctionne comme un extracteur de caractéristiques des images. Une image est passée à travers une succession de filtres, ou noyaux de convolution, créant de nouvelles images appelées cartes de convolutions ou carte de caractéristiques (feature map). Certains filtres intermédiaires, de sous-échantillonnage, réduisent la résolution de l'image par une opération de maximum local. Au final, les cartes de convolutions sont mises à plat (opération flatten) et concaténées en un vecteur de caractéristiques, appelé code CNN. Ce code CNN en sortie de la partie convolutive peut être utilisée comme vecteur de caractéristiques pour classer l ?image. En pratique, le réseau CNN est tronqué pour ne garder que la partie convolutive, comme illustré dans la figure 2.11. Cette partie est dite gelée. Ce réseau prend en entrée une image et produit en sortie le code CNN. Chaque image est ainsi transformée en un vecteur de caractéristiques, qui est utilisé pour entraîner un nouveau classifieur (svm, knn ou autres) afin de catégoriser l'image. Cette

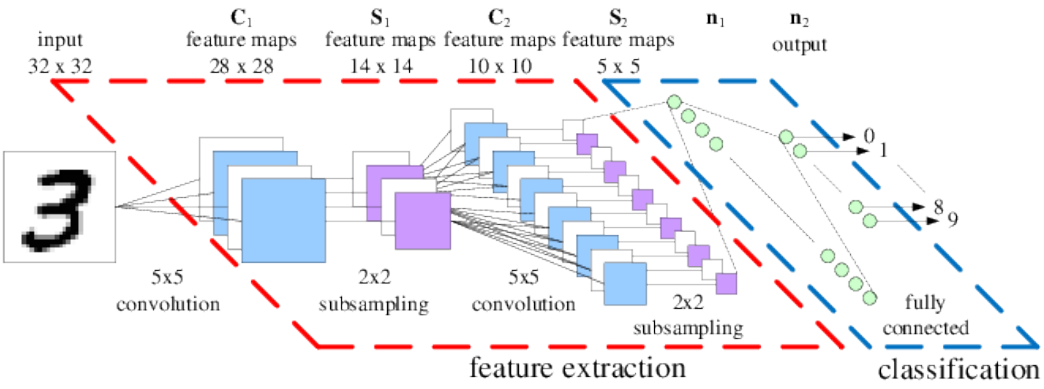

FIGURE 2.11 – Utilisation d'un réseau de neurones convolutif pré-entraîné pour l'extraction des caractéristiques [2].

méthode présente de nombreux intérêts pratiques. Le premier intérêt est la réduction de la dimension du problème car l'image est transformée en un vecteur de petite dimension, appelé "code CNN", qui extrait des caractéristiques généralement très pertinentes. Par ailleurs, il serait possible de tester différents classificateurs qui prennent comme entrée le code CNN.

2.4.4.2 Fine Tuning

Dans la méthode Fine tuning, la connaissance contenue dans le réseau convolutif pré-entraîné peut être utilisée comme initialisation du modèle. Par la suite, le modèle est ré-entraîné de manière plus fine pour s'adapter au nouveau problème de classification. Cette méthode permet de se servir d'un modèle pré-entraîné comme initialisation pour l'entraînement d'un nouveau problème similaire [112]. Par exemple dans notre travail, le problème à traiter est la reconnaissance de scènes et la reconnaissance d'objets. Le *Fine tuning* sur des images consiste en quelques sorte à prendre un système visuel déjà bien entraîné sur une tâche de classification particulière pour le raffiner sur une tâche similaire. L'entraînement dans le *Fine tuning* consiste à adapter la dernière couche pour correspondre au nombre de catégories du nouveau problème. Dans cette procédure, la dernière couche joue le rôle d'un classifieur final avec un perception multi-couches pré-initialisé. Il est possible aussi d'ajouter des couche à l'architecture pré-existante, ou de modifier la taille des couches préalables.

Des équipes de recherche se spécialisent dans l'amélioration des CNNs. Elles publient leurs innovations techniques, ainsi que le détail des réseaux entraînés sur des bases de données de références. Le challenge ImageNet (ILSVRC) [1] fournit par exemple 1.2 millions d'images classées en 1000 catégories. Ces CNNs entraînés sont disponibles dans plusieurs librairies entre autres Keras [2] et Caffe [3] ou ils sont regroupés dans des dépôts GitHub pour d'autres librairies.

1. http://www.image-net.org/challenges/LSVRC/
2. https://github.com/teammcr192/deep-CNN-keras
3. https://github.com/BVLC/caffe/tree/master/models

2.5 Sélection de caractéristiques pour la classification

La sélection de caractéristiques permet d'en choisir les plus pertinentes, celles adaptées à la résolution d'un problème particulier afin d'améliorer les performances globales du système. La présence des caractéristiques redondantes ou inutiles dans le modèle peut conduire à une détérioration des performances en généralisation [113]. La sélection de caractéristiques est définie comme un processus de recherche permettant de trouver un sous-ensemble "pertinent" de caractéristiques parmi celles de l'ensemble de départ. La notion de pertinence d'un sous-ensemble de caractéristiques dépend toujours des objectifs et des critères du système [114].
Dans le cadre de l'apprentissage supervisé pour l'annotation des images, le principe est de sélectionner le sous-ensemble de caractéristiques permettant de discriminer au mieux les différentes classes de données.

2.5.1 Pertinence de caractéristiques

La performance d'un algorithme d'apprentissage dépend fortement des caractéristiques utilisées dans la tâche d'apprentissage. La performance de cette tâche peut être détériorée par la présence de caractéristiques redondantes ou non pertinentes. Il existe plusieurs définitions de la pertinence d'une caractéristique dans la littérature. Selon Jhon et al. [115], une caractéristique peut être classée comme suit :

- **Très pertinente** : une caractéristique est considérée comme très pertinente si son absence entraîne une dégradation significative de la performance du système de classification.

- **Peu pertinente** : une caractéristique X_i est considérée comme peu pertinente si elle n'est pas "très pertinente" et s'il existe un sous-ensemble S tel que la performance de $S \cup X_i$ est meilleure que la performance de S. Une caractéristique peu pertinente peut contribuer parfois à la précision de la prédiction.

- **Non pertinente** : une caractéristique non pertinente ne peut jamais contribuer à la précision de la prédiction. Une caractéristique est dite pertinente si elle est très ou peu pertinente, sinon elle est non pertinente.

Dans ce qui suit, nous allons voir les méthodes utilisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection. Elles peuvent être classées en trois catégories principales [116] : *les méthodes par filtre*, *les méthodes encapsulante* et *les méthodes intégrées*.

2.5.2 Méthode par filtre (Filter)

Dans la méthode "*filter*", un critère d'évaluation est utilisé pour déterminer la pertinence d'une caractéristique selon des mesures qui reposent sur des données d'apprentissage. Parmi les exemples de critères utilisés pour la sélection d'un sous-ensemble de caractéristique sont "la corrélation" et "l'information mutuelle" entre ces caractéristiques. Le sous-ensemble qui optimise le critère d'évaluation est sélectionné. Cette méthode est considérée comme une étape de prétraitement (filtrage). L'évaluation se fait généralement indépendamment d'un classifieur [115]. La figure 2.12 présente la procédure de la méthode "*filter*". Le grand avantage de ces méthodes de filtrage est leur efficacité calculatoire et leur robustesse face au sur-apprentissage. Cependant, ces méthodes ne prennent pas en considération les interactions entre caractéristiques et tendent à sélectionner des caractéristiques comportant de l'information redondante plutôt que complémentaire [117].

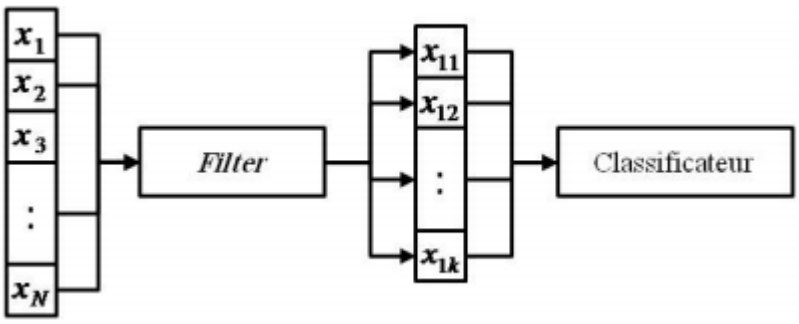

FIGURE 2.12 – Procédure de la méthode Filter [114].

2.5.3 Méthodes enveloppantes (Wrapper)

Le principal inconvénient des approches "*filter*" est d'ignorer l'influence des caractéristiques sélectionnées sur la performance du classificateur à utiliser

par la suite. Le concept "*wrapper*" pour la sélection des caractéristiques a été introduit par Kohavi et John [118] pour résoudre ce problème. La méthode enveloppante, appelée aussi "*wrapper*", teste différents sous-ensembles de jeux de caractéristiques et conserve au final le sous-ensemble qui donne les meilleurs performances. L'évaluation se fait à l'aide d'un classificateur qui estime la pertinence d'un sous-ensemble donné de caractéristiques. Le sous-ensemble de caractéristiques retenu par cette méthode est bien adaptés à l'algorithme de classification utilisé, toutefois, il n'est pas forcément valide pour un autre type de classifieur. La figure 2.13 illustre la procédure de la méthode "*wrapper*".

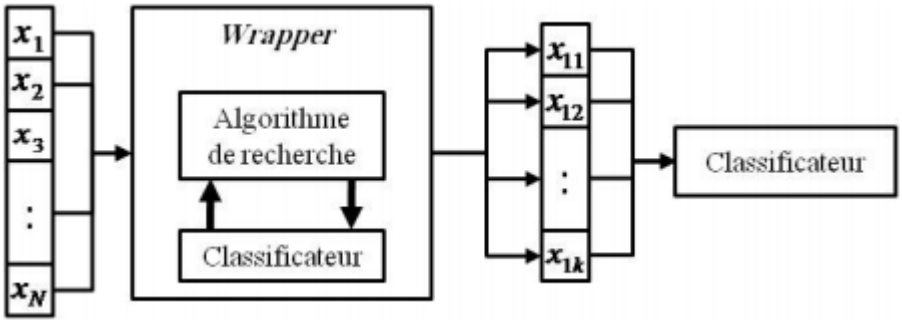

FIGURE 2.13 – Procédure de la méthode Wrapper [114].

L'inconvénient majeur de cette méthode est son coût en terme de temps de calcul puisque elle relance plusieurs fois l'algorithme d'apprentissage avec chaque sous-ensemble de caractéristiques possibles afin de retenir le meilleur. Le problème de la complexité de cette technique rend impossible l'utilisation d'une stratégie de recherche exhaustive (problème NP-complet) [116]. Par conséquent, des méthodes de recherche heuristiques ou aléatoires peuvent être utilisées. La deuxième limitation de la méthode "*wrapper*" est que l'évaluation de caractéristiques se fait par un seul classificateur lors la sélection. Ainsi le sous-ensemble sélectionné dépend toujours du classificateur utilisé [116].

2.5.4 Méthode intégrée (Embedded)

A la différence des méthodes "*wrapper*" et "*filter*", les méthodes "*intégrées*", appelées aussi "*embedded*", introduisent la sélection des caractéristiques lors de l'apprentissage [119]. Ce mécanisme d'intégration de la sélection

des caractéristiques dans le processus d'apprentissage est utilisé par les algorithmes de type SVM, Adaboost [120] et les arbres de décisions [114]. Dans les méthodes "*wrapper*", la base d'apprentissage est séparée en deux : une base d'apprentissage et une base de validation pour valider le sous-ensemble de caractéristiques sélectionné. Dans les méthodes "*embedded*", la base d'apprentissage entière peut être utilisée pour établir le système. Par ailleurs, ces méthodes sont plus rapide que les méthodes "*wrapper*" vu qu'elles évitent que le classificateur recommence à zéro pour chaque sous-ensemble de caractéristiques.

2.6 Conclusion

Nous avons présenté dans ce chapitre les notions liées à l'apprentissage automatique. Nous avons vu la définition du concept de l'apprentissage automatique et ses différents types : supervisé et non supervisé. Nous avons aussi évoqué en détails quelques méthodes génératives pour l'apprentissage supervisé. Nous avons présenté plusieurs méthodes et algorithmes usuels en apprentissage discriminatif. Enfin, nous avons vu comment améliorer les performances d'un algorithme d'apprentissage en utilisant la sélection des caractéristiques qui permet de choisir les meilleurs descripteurs pour apprendre un modèle.

Dans le chapitre suivant, nous nous intéressons aux méthodes d'apprentissage génératives. Nous nous focalisons sur les modèles graphiques probabilistes et plus particulièrement les réseaux bayésiens.

Chapitre 3
LES RÉSEAUX BAYÉSIENS

3.1 Introduction

Ce chapitre bibliographique présente une approche générative pour l'apprentissage supervisé, et plus précisément les réseaux bayésiens. Les approches génératives permettent de modéliser la structure d'un système. La réponse au besoin de l'utilisateur est inférée de ce modèle. Contrairement aux méthodes discriminatives qui ne s'intéressent qu'aux relations entrées-sorties, ces approches modélisent en plus les liens qui existent au sein des variables d'entrée [121]. Ainsi, le travail de modélisation est plus important. Le terme "*génératif*" désigne le fait que la règle de décision, déduite d'après les relations de Bayes [122], soit basée sur une modélisation de la probabilité $p(X|Y)$ qui "*génère*" les observations X pour une classe Y donnée. Lorsque le protocole expérimental le permet, ces méthodes peuvent aussi facilement tenir compte des probabilités a priori $p(Y)$ d'apparition de chaque classe. En cas d'ignorance, il est d'usage de considérer les classes comme équiprobables.

L'approche générative regroupe une vaste catégorie de classifieurs, dont le plus utilisé est le réseau bayésien [122]. Les réseaux bayésiens sont des modèles graphiques probabilistes et qui permettent de modéliser les dépendances entres les différents composants d'un système simple ou complexe [123, 124]. Ils reposent sur la théorie de graphe [125] et la théorie des probabilités [126]. Ils sont un moyen simple de représenter une distribution de probabilité jointe d'un ensemble de variables aléatoires, de visualiser les propriétés de dépendance conditionnelle, et ils permettent d'effectuer des calculs complexes comme l'apprentissage des probabilités et l'inférence, avec des manipulations graphiques. Les réseaux bayésiens ont été initiés par *Judea Pearl* dans les années 80 [127]. Ils se sont révélés des outils très efficaces et pratiques pour la représentation de connaissances. Un réseau bayésien semble donc approprié pour représenter et classer les albums dans des évènements.

Pour ces raisons, nous nous intéressons aux réseaux bayésiens dans le cadre de cette thèse. Au cours de cette section, nous focalisons notre étude sur les

réseaux bayésiens afin de fournir au lecteur un ensemble d'éléments qui lui permet d'appréhender plus facilement la suite de cette thèse.

3.2 Définition des réseaux bayésiens

Il est possible de rencontrer dans la littérature différentes dénominations pour ces réseaux, telles que les réseaux de croyance (belief networks), les modèles graphiques probabilistes orientés (probabilistic oriented graphical models), les réseaux probabilistes (probabilistic networks), ou encore les réseaux d'indépendances probabilistes (probabilistic independence networks) [127], [128], [123], [129].

Ces réseaux permettent de représenter de manière concise la distribution de probabilité jointe sur un ensemble de variables aléatoires. La structure de ce type de réseau est simple : un graphe dans lequel les nœuds représentent les variables aléatoires, et les arcs reliant ces dernières sont rattachées à des probabilités conditionnelles. Notons que le graphe est acyclique, c'est-à-dire il ne contient pas de boucle.
Plus formellement, un réseau bayésien $B = \{G, \Theta\}$ est défini comme suit [130]

- un graphe dirigé sans circuit $G = (X, E)$ dont les sommets (nœuds) sont associés à un ensemble de variables aléatoires $X = \{X_1, ..., X_n\}$, et E est l'ensemble des arcs,

- un ensemble des probabilités $\theta_i = \{p(X_i|pa(X_i))\}$ de chaque nœud X_i conditionnellement à l'état de ses parents $pa(X_i)$ dans G.

Un réseau bayésien B représente une distribution de probabilités sur X dont la loi jointe peut se simplifier de la manière suivante :

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i|pa(X_i)) = \prod_{i=1}^{n} \theta_i \tag{3.1}$$

Cette décomposition d'une fonction globale en un produit de termes locaux dépendant uniquement du nœud considéré et de ses parents dans le graphe, représente une propriété fondamentale des réseaux bayésiens.

3.3 Indépendance conditionnelle et d-séparation

3.3.1 Indépendance conditionnelle

Soient X et Y deux variables représentées dans un graphe G. Les variables X et Y sont inconditionnellement indépendants, si l'observation de la variable Z rend la variable X indépendante de Y. Cela est noté : $X \perp Y|Z$. Ce qui revient à dire en terme de probabilité que [129] :

$$p(X|Y,Z) = p(X|Z) \tag{3.2}$$

La figure 3.1 (b) illustre un graphe correspondant à un réseau bayésien. Dans ce graphe, les variables X et Y sont indépendants étant donné Z.

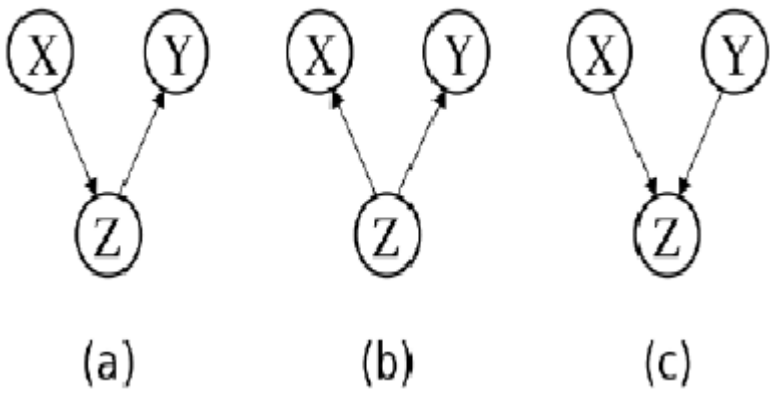

FIGURE 3.1 – Connexion en série (a), divergence (b), ou convergente (c) [129].

Les connexions entre les nœuds définissent les lois de circulation de l'information dans le graphe. Patrick Naim distingue trois types de connexions [129] :

- connexion en série (figure 3.1 a) : l'information ne peut circuler entre X et Y que si la valeur de Z n'est pas connue, sinon c'est directement la connaissance sur le nœud Z qui influe ;

- connexion divergente (figure 3.1 b) : l'information ne peut circuler entre X et Y que si la valeur de Z n'est pas connu ;

- connexion convergente ou connexion en V (figure 3.1 c) : l'information ne peut circuler entre X et Y que si la valeur de Z est connue.

3.3.2 d-Séparation

La notion de d-séparation est essentielle pour le calcul des probabilités *a posteriori*. Elle permet de définir l'indépendance conditionnelle entre certains nœuds. Dans un réseau bayésien, deux nœuds X et Y sont d-séparés par un ensemble de variables Z si tous les chemins (sans tenir comptes de l'orientation) entre X et Y sont bloqués (en tenant compte de l'orientation). Un chemin est bloqué s'il empêche X et Y d'être dépendants. En d'autres termes, X et Y sont d-séparés par Z si pour tous les chemins entre X et Y, l'une au moins des deux conditions suivantes est vérifiée [131] :

- Le chemin converge en un nœud W, tel que $W \neq Z$, et W n'est pas une cause directe de Z.
- Le chemin passe par Z, est soit divergent, soit en série au nœud Z.

X est d-séparé de Y par Z est noté $< X|Z|Y >$.
La figure 3.2 illustre un exemple de d-séparation. Dans le graphe, A et D sont d-séparés par B car le chemin orienté allant de A vers D passe par B. De même, le nœud D d-sépare les nœuds ABC des nœuds EFG.

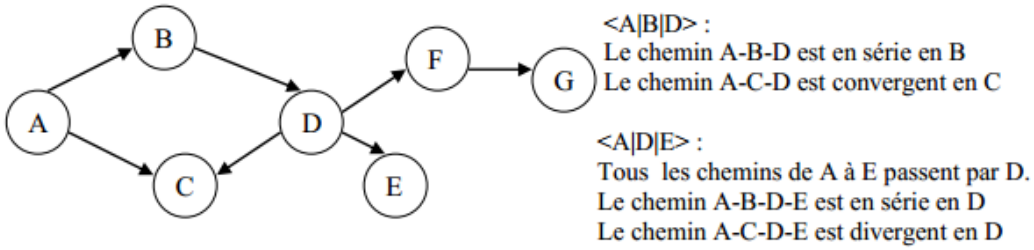

FIGURE 3.2 – Exemple de d-séparation [132].

3.4 Construction d'un réseau bayésien

La construction d'un réseau bayésien sachant un ensemble d'observations peut se décomposer en deux étapes. La première consiste à apprendre sa struc-

ture, c'est-à-dire le graphe associé au réseau. Une fois sa structure connue, il est nécessaire d'estimer ses distributions de probabilités conditionnelles. Nous nous restreignons ici au cas des réseaux bayésiens discrets, où les distributions de probabilités sont décrites par un ensemble de paramètres. De plus, nous supposons que les observations sont complètes, nous ne traiterons donc pas de situation où les données sont manquantes.

3.4.1 Apprentissage de structure

Dans la plupart des applications, cette étape s'effectue manuellement à l'aide d'experts. Dans ce cas, des itérations sont souvent nécessaires pour aboutir à une description consensuelle des itérations entre les variables observées. Les spécialistes en ingénierie de la connaissance interrogent les experts et ajoutent les nœuds et les liens au réseau bayésien sur la base de la connaissance recueillie. Il est possible d'apprendre la structure d'un réseau bayésien de manière automatique en recherchant la carte parfaite des distributions de probabilités conditionnelles sous-jacentes aux observations.

L'apprentissage automatique de la structure permet de déterminer la structure optimale d'un réseau bayésien à partir de l'information contenue dans les données observées. Trois grandes familles d'approches d'apprentissage automatique de structure sont distinguées [133] :

- **Les méthodes de recherche d'indépendances** : Les algorithmes tels que IC (Inductive causality) [134] , PC [135], FCI [136] sont des exemples connus pour ce type de méthodes. Ces méthodes consistent à déterminer dans un premier temps un graphe non orienté qui représente les différentes indépendances conditionnelles existantes entre les variables observées à l'aide du test d'indépendances conditionnelles existantes entre les variables observées, puis à orienter ce graphe pour obtenir un réseau bayésien. Ces algorithmes sont peu efficaces dans le cas de problèmes de grande taille puisque la détermination de ces indépendances est exponentielle en fonction du nombre de variables. Les méthodes de recherche d'indépendances conditionnelles ont donc une complexité élevée.

- **Les méthodes à base de scores** : La deuxième approche consiste à parcourir tous les graphes possibles, associer un score à chaque graphe,

puis choisir le graphe ayant le score le plus élevé. En général les scores existants sont tous des approximations différentes de la vraisemblance marginale [137].Parmi les fonctions de score existantes, nous citons les plus courantes[129] : le score BD (Bayesian Dirichlet) et le score BDe (Bayesian Dirichlet Equivalent [138]). Par exemple, dans le score BD [139], en partant d'une loi a priori sur les structures possibles $p(G)$, le but est d'exprimer la probabilité a posteriori des structures possibles sachant que les données D ont été observées $p(G/D)$ ou plus simplement $p(G, D)$ tel que :

$$ScoreBD(G, D) = p(G, D) = \int_{\Theta} L(D|\theta, G)p(\theta|G)p(G)\mathrm{d}\theta \tag{3.3}$$

$$= p(G) \int_{\Theta} L(D|\theta, G)p(\theta|G)\mathrm{d}\theta \tag{3.4}$$

$$\tag{3.5}$$

où $L(D|\theta, G)$ désigne la fonction de vraisemblance et θ désigne les paramètres du modèle tel que $\theta \in \Theta$. Toutefois, les méthodes à base de score ne sont pas simples, principalement à cause de la taille super-exponentielle de l'espace de recherche en fonction du nombre de variables.

- **Les méthodes hybrides** : Le but de ces méthodes est de combiner les avantages des méthodes de recherche d'indépendances et celles basées sur les scores. Inspirées par le principe "diviser pour régner", les méthodes hybrides consistent en deux étapes :

  - une recherche locale, qui permet d'obtenir un voisinage contenant toutes les dépendances (indépendances) locales intéressantes avec l'aide de tests d'indépendances. Citons par exemple PCMB (parents and children based Markov boundary algorithm) [140], MBOR (Markov Boundary search using the OR condition) [141].
  - une optimisation globale, qui permet de faire une recherche sur l'espace des DAGs (Directed Acyclic Graph) en se restreignant aux dépendances locales trouvées précédemment par exemple MMHC(Max-Min Hill-Climbing) [142].

Quelle que soit l'approche choisie, l'apprentissage automatique de la structure de réseau bayésien à partir de données pose quelques problème communs :

- La recherche de la structure optimale reste un problème NP-difficile [143]. Ces méthodes ne sont pas si simples, principalement à cause de la taille super-exponentielle de l'espace de recherche en fonction du nombre de variables.

- L'apprentissage de la structure est grande consommatrice de données [144]. C'est-à-dire qu'il faut des bases de taille importante afin de pouvoir obtenir un modèle de qualité. Cette limite devient un problème crucial dans les domaines expérimentaux où le recueil de données est difficile, cher, etc.

Donc, un véritable défi est alors d'abord d'optimiser la recherche (soit transposer le problème en un espace plus petit, soit utiliser une stratégie de recherche optimisée) en travaillant sur un ensemble d'apprentissage réduit.

3.4.2 Apprentissage des paramètres

L'apprentissage des paramètres pour une structure particulière consiste à estimer les distributions de probabilités *a priori* ou les paramètres des lois de probabilités à partir des données disponibles. Dans le cas où l'on dispose de données complètes, on peut utiliser le maximum de vraisemblance qui utilise la fréquence d'apparition d'un événement dans les données [145] comme nous allons le détailler dans la section suivante. Dans le cas échéant, si les données sont incomplètes, l'algorithme Expectation-Maximisation peut être utilisé [146].
Dans cette thèse, nous nous intéressons au cas des données complètes et discrètes. Dans le cas discret, un paramètre $\theta_i = p(X_i|pa(X_i))$ est un tableau contenant l'ensemble de probabilités de la variables X_i pour chacune de ses valeurs possibles sachant chacune des valeurs prises par l'ensemble de ses parents $pa(X_i)$ dans le graphe G. L'estimation de ces paramètres sera détaillée dans la section suivante.

3.5 Inférence Bayésienne

Dans un premier temps, nous allons présenter le théorème de Bayes et l'algorithme de maximum de vraisemblance (MV). Par la suite, nous verrons les deux principales caractéristiques de formalisme bayésien, c'est à dire l'utilisation des connaissances a priori avant la prise en compte des observations, et la capacité d'éliminer les paramètres du modèle qui ne sont pas représentatifs du problème par marginalisation. Pour obtenir la probabilité *a posteriori* des variables cibles, la connaissance et les observations sont combinées par l'estimateur du maximum *a posteriori* (MAP). Ensuite, nous présentons les modèles paramétriques les plus courants.

3.5.1 Théorème de Bayes

Soient A et B deux événements aléatoires tels que $p(B) \neq 0$. La probabilité de A, conditionnellement à la réalisation de B, est par définition exprimée par la relation suivante [147] :

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{3.6}$$

où $p(A,B)$ est la probabilité que deux événement A et B aient lieu simultanément. Puisque $p(A,B) = p(B,A)$, alors les deux probabilités conditionnelles $p(A|B)$ et $p(B|A)$ sont reliées par :

$$p(A|B) = \frac{p(A)p(B|A))}{p(B)} \tag{3.7}$$

Cette équation est une conséquence triviale de la définition de la probabilité conditionnelle, et appelé *Formule de Bayes* (ou Théorème de Bayes) [148].

Chaque terme du théorème de Bayes a une dénomination usuelle. Le terme $p(A)$ est la probabilité *a priori* de A. L'appellation *a priori* exprime le fait qu'elle a été établie agréablement à l'observation des données A. Le terme $p(A|B)$ est appelée la probabilité *a posteriori* de A sachant B (ou encore de A sous condition B). L'inférence est fondée sur la distribution *a posteriori*. Le terme $p(B|A)$, pour un B connu, est appelée *la fonction de vraisemblance* de A. De même, le terme $p(B)$ est appelé la probabilité *marginale* de B. Comme

le dénominateur est indépendant de A, c'est uniquement une constante de normalisation, la formule de Bayes peut s'écrire comme suit :

$$p(A|B) \propto p(A)p(B|A) \tag{3.8}$$

Dans un modèle statistique, le passage de la distribution *a priori* à la distribution *a posteriori* des paramètres du modèle, exprimé par la formule de Bayes, peut être interprété comme une mise à jour de la connaissance, sur la base des observations.

3.5.2 Maximum de Vraisemblance (MV)

Les méthodes d'estimation du maximum de vraisemblance cherchent à trouver une valeur estimée de $\hat{\theta}$ pour un paramètre θ à partir d'un ensemble d'échantillons donnée [149]. Elles possèdent de bonnes propriétés de convergence et d'efficacité quand le nombre d'échantillon est très grand [150]. Soit X une variable aléatoire de densité de probabilité $f(x|\theta)$. Soit $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ un échantillon de données indépendantes et identiquement distribuées.*iid* généré par la densité de probabilité $f(x|\theta)$ avec les paramètres $\theta \in \Theta$. Dans le cas où les variables aléatoires sont indépendantes, la fonction de vraisemblance associée L (vraisemblance de θ par rapport à $\mathcal{D}$) est définie sur Θ par le produit de densités de probabilité individuelles exprimée par l'équation 3.9 [149] :

$$L(\theta) = f(D|\theta) = \prod_{j=1}^{n} f(x_j|\theta) \tag{3.9}$$

La méthode du maximum de vraisemblance consiste à trouver les valeurs $\hat{\theta}$ de θ qui maximisent la vraisemblance $L(\theta)$, tel que $\hat{\theta} = max_{\theta \in \Theta} L(\theta)$. Cette estimation correspond à la valeur de θ qui supporte le mieux l'ensemble de données observée.

Grâce à la monotonie de la fonction *Logarithmique*, $\hat{\theta}$ peut être trouvé en maximisant le logarithme de la fonction de vraisemblance $l(\theta)$ (appelé fonction *log-vraisemblance*) [149]. Cela possède l'avantage en calcul de remplacer un produit par une somme.

$$l(\theta) = log(L(\theta)) = \sum_{j=1}^{n} log(f(x_j|\theta)) \tag{3.10}$$

On peut écrire la solution sous la forme :

$$\hat{\theta}_{MV} = arg \max_{\theta \in \Theta} l(\theta) \tag{3.11}$$

L'avantage de l'estimation de maximum de vraisemblance est la simplicité et la rapidité de son calcul [150]. Le principe du maximum de vraisemblance fournit une approche d'estimation bien connue comme dans le cas de distributions normales et plusieurs autres distributions tel que détaillé dans la section 3.6. De point de vue interprétabilité, le maximum de vraisemblance est facile à comprendre et à interpréter du moment où cette méthode retourne un seul modèle optimal pour l'ensemble des échantillons disponibles.

3.5.3 Les lois a priori

L'une des caractéristiques fondamentales de l'analyse bayésienne est l'utilisation de la densité de probabilité décrivant l'état de connaissance ou d'ignorance concernant les paramètres avant la prise en compte des observations [151]. Le choix de cette densité qui est dite *a priori* est le problème le plus délicat et le plus critique de l'analyse bayésienne. En effet, il est très rare que l'information a priori soit suffisante pour pouvoir déterminer cette densité donc il faut faire des approximations. Le choix des lois *a priori* est une étape fondamentale dans l'analyse bayésienne [151]. Ce choix peut avoir différentes motivations. Les stratégies sont diverses. Elles peuvent se baser sur des expériences du passé ou sur une intuition, une idée que le praticien a sur le phénomène aléatoire qu'il est en train de suivre. Elles peuvent être également motivées par des aspects calculabilité. Dans notre cas, nous ne présentons que les deux types de densités *a priori* les plus courantes : la *densité a priori conjuguée* et la *densité a priori non informative*.

3.5.3.1 Lois a priori conjuguées

La *loi a priori conjuguée*, proposée par Raiffa et Schlaider [152] , peut être considérée comme un point de départ pour l'élaboration de distributions

Vraisemblance	a priori	a posteriori
Binomial $B(n,\theta)$	Beta $Be(\alpha,\beta)$	Beta $Be(\alpha+x,\beta+n-x)$
Multinomial $M_k(\theta_1,...,\theta_k)$	Dirichlet $D(\alpha_1,...,\alpha_k)$	Dirichlet $D(\alpha_1+x_1,...,\alpha_k+x_k)$
Gamma $G(v,\theta)$	Gamma $G(\alpha,\beta)$	Beta $Be(\alpha+v,\beta+x)$
Poisson $P(\theta)$	Gamma $G(\alpha,\beta)$	Gamma $Be(\alpha+x,\beta+1)$

TABLE 3.1 – Exemples de distributions a priori conjuguées [154].

a priori fondées sur des informations a priori limitées.

Définition [152] :
La loi des observations étant supposée connue, on se donne une famille F de lois de probabilité sur Θ. On suppose que la loi a priori appartient à F. Si dans ces conditions, la loi a posteriori appartient encore à F, on dit que la loi a priori est conjuguée.

L'avantage des familles conjuguées est de simplifier les calculs [153]. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs. Le tableau récapitulatif 3.1 présente les distributions conjuguées associées à un certain nombre de modèles classiques utilisables pour n-échantillons indépendants.

Il est intéressant d'utiliser des familles de densités *a priori conjuguées* telles que celles-ci aient la même forme fonctionnelle que la fonction de vraisemblance. Dans ce cas, le passage de la fonction de vraisemblance à la densité a posteriori se réduit à un changement de paramètre et non à une modification de la forme fonctionnelle de la famille correspondante.

3.5.3.2 Lois a priori non informatives

Le cas où aucune connaissance a priori n'est disponible, on utilise des densités a priori qui sont dites non informatives. Nous présentons deux exemples de lois a priori non informatives : *loi a priori uniforme* et *loi a priori de Jeffrey*.

Loi a priori uniforme
La densité *a priori non informative* la plus simple et la plus communément utilisée est la densité uniforme [155], ce choix repose sur l'équiprobabilité des valeurs possibles du paramètre sur son domaine de définition et, donc,

n'apporte aucune information supplémentaire sur θ. Cette densité a priori est définie par : $p(\theta) = k$ où k est une constante.

Loi a priori de Jeffrey [156]

Jeffrey (1961) a proposé une loi a priori en se basant sur le principe d'invariance par transformation [157]. Par définition, un estimateur est invariant si, pour une reparamétrisation $\phi = f(\theta)$ du modèle, les estimations $\hat{\theta}$ et $\hat{\phi}$ sont telles que $\hat{\phi} = f(\hat{\theta})$.

Soit $p(y|\theta)$ la fonction de vraisemblance et $\eta = f(\theta)$, telle que f est une transformation bijective de θ telle que $\eta_i = f_i(\theta), i = 1, ..., d$, alors les densité *a priori non informative de Jeffrey* est telle que :

$$|J(\theta)|^{\frac{1}{2}} d\theta = |J(\eta)|^{\frac{1}{2}} d\eta \tag{3.12}$$

Où $|J(.)|$ est le déterminant de la matrice d'information de Fisher définie par :

$$J_{ij}(\theta) = -E\left[\frac{\partial^2 lnp(y|\theta)}{\partial\theta_i \partial\theta_j}\right] \tag{3.13}$$

Où E est l'espérance. Si les paramètres $\theta_i = 1, ..., d$ sont a priori indépendants, la densité a priori de Jeffrey de θ devient :

$$p(\theta) = \prod_{i=1}^{d} p(\theta_i) \propto \left[\prod_{i=1}^{d} J(\theta_i)\right]^{\frac{1}{2}} \tag{3.14}$$

Où $J(\theta_i)$ est la fonction d'information de Fisher de θ_i.

3.5.4 Marginalisation

La deuxième caractéristique du paradigme bayésien est la capacité d'éliminer des paramètres qui ne nous intéressent pas par *marginalisation* [158]. Supposons que le vecteur des paramètres θ puisse se décomposer en deux vecteurs tels que $\theta = (\theta_1, \theta_2) \in \Theta 1 \times \Theta_2$ et que seul θ_1 nous intéresse. La *densité a posteriori marginale* de $\theta_1, p(\theta_1|Y)$ est déduite de la densité a posteriori conjointe $p(\theta|Y)$ par intégration de θ_2 sur son domaine de définition Θ_2.

$$p(\theta_1|Y) = \int_{\Theta 2} p(\theta|Y) d\theta_2 \tag{3.15}$$

Cette technique implique une réduction de la dimension de l'espace de paramètres à estimer qui peut être très intéressante lors de l'analyse bayésienne.

3.5.5 Théorie de la décision : Estimateur du maximum a posteriori (MAP)

Dans certains cas, un expert du domaine peut avoir des connaissances *a priori* sur certaines variables ou même sur le problème en entier. Ces informations sont très utiles et doivent être exploitées et introduites en tant qu'*a priori* sur les paramètres θ, surtout dans le cas où il n'y a que très peu de données. Les informations *a priori* n'indiquent pas exactement les valeurs des paramètres, l'incertain sera alors modélisé en considérant θ comme variable aléatoire et en définissant une densité *a priori* $f(\theta)$ pour cette variable.

La densité a priori, $f(\theta)$, nous indique la valeur que doit avoir θ avant de voir les données. On combine ensuite cette information *a priori* avec les informations issues des données, notamment la densité obtenue par maximum de vraisemblance $f(D|\theta)$, en utilisant la règle de Bayes. On obtient ainsi une densité a posteriori de θ [159] :

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{f(D)} = \frac{(D|\theta)f(\theta)}{\int_{\theta' \in \Theta} f(D|\theta')f(\theta')d\theta'} \tag{3.16}$$

L'estimation de la densité en un point X est donnée par l'équation suivante :

$$f(X|D) = \int_{\Theta} f(X|\theta)f(\theta|D)d\theta \tag{3.17}$$

L'estimation bayésienne est alors une moyenne des prédictions en utilisant toutes les valeurs possibles de θ, pondérées par leur probabilité. Le calcul de cette intégrale de l'équation 3.17 peut être un problème difficile à résoudre. Dans le cas où ce calcul n'est pas faisable, on peut approcher l'intégralité par un seul point. L'estimation du maximum a posteriori (MAP) peut rendre le calcul facile en faisant l'hypothèse que $f(\theta|D)$ présente un pic [160]

$$\hat{\theta}_{MAP} = arg\max_{\theta} f(\theta|D) \tag{3.18}$$

Par la suite, la densité dans l'équation 3.18 est remplacée par un seul point comme suit :

$$f(X|D) = f(X|\hat{\theta}_{MAP}) \tag{3.19}$$

Le choix de la loi a priori est une difficulté majeure de l'approche bayésienne. Il est généralement guidé par les connaissances a priori que l'on peut avoir du domaine étudié. Lorsque la connaissance a priori est très précise elle peut impliquer l'utilisation d'une loi a priori entièrement spécifiée. On peut également utiliser une loi a priori non informative telle que la loi uniforme, en l'absence d'information a priori.

Il est possible d'avoir recours à la famille des lois a priori conjuguées dont l'intérêt est d'assurer le calcul analytique de la densité a posteriori. En effet, elles sont définies de telle sorte que la loi a posteriori appartient à la même famille de lois que la loi a priori.

3.6 Exemples de modèles usuels

3.6.1 Loi de Bernoulli

Forme de la loi [161]

Soit une variable aléatoire X discrète qui suit une loi de Bernoulli de paramètre θ. Dans une loi de Bernoulli, il existe deux sorties possibles : un événement se produit ou non. Si l'évènement se produit, la variable aléatoire de Bernoulli X prend la valeur 1 avec une probabilité θ, si l'événement ne se produit pas, la probabilité de l'événement est $1-\theta$ et la variable aléatoire X prend la valeur 0. Cela est défini par :

$$p(X) = \theta^X(1-\theta)^{1-X}, X \in \{0,1\} \tag{3.20}$$

Estimation par Maximum de Vraisemblance [162]

Soit un échantillon de *variables indépendantes et identiquement distribuée.iid* $\mathcal{D} = \{X_t\}_{t=1}^N$ où $X_t \in \{0,1\}$, pour calculer l'estimateur de maximum de vraisemblance $\hat{\theta}$, nous définissons la fonction $log-vraisemblance$ par l'équation :

$$l(\theta) = log\prod_{t=1}^{N}\theta^{X_t}(1-\theta)^{1-X_t} = \sum_t X_t log\theta + \left(N - \sum_t X_t\right) log(1-\theta) \tag{3.21}$$

$\hat{\theta}$ qui maximise le $log - vraisemblance$ peut être calculé en résolvant $dl/d\theta = 0$.

$$\hat{\theta} = \frac{\sum_t x_t}{N} \tag{3.22}$$

L'estimateur de θ est donné par le nombre d'occurrence de l'événement divisé par le nombre total d'individus.

Estimateur par MAP [162]

La distribution a priori conjuguée pour une loi de Bernoulli est une distribution $Beta$ de paramètres α et β, notée $Beta(\alpha, \beta)$

Soit un échantillon iid $D = \{X_t\}_{t=1}^N$ où $X_t \in \{0, 1\}$, et une distribution a priori de type Beta, notée par $Beta(\alpha, \beta)$ définie par :

$$Beta(\theta, \alpha, \beta) = \gamma\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{3.23}$$

avec $\gamma = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ est une constante qui ne dépend pas de θ.

Pour estimer le Maximum a posteriori de θ, il faut maximiser la fonction

$$f(x|\theta)Beta(\theta, \alpha, \beta) = \theta^x(1-\theta)^{1-x}\gamma\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{3.24}$$

L'estimateur MAP $\hat{\theta}_{MAP}$ de θ est donné par

$$\hat{\theta}_{MAP} = \frac{\sum_t x_t + \alpha - 1}{N + \alpha + \beta - 1} \tag{3.25}$$

Une expérience qui consiste à répéter $nfois$ de manière indépendante l'épreuve de Bernoulli de paramètre θ est appelée un schéma de Bernoulli d'ordre n. La loi Binomiale de paramètres n et θ est la loi de probabilité de la variable aléatoire X prenant comme valeurs le nombre de succès obtenus au cours des n épreuves du schéma de Bernoulli. Il est donc possible d'obtenir la probabilité de k succès dans une répétition de n expériences par :

$$p(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k(1-\theta)^{n-k} \tag{3.26}$$

3.6.2 Loi Multinomiale [161]

Une variable X discrète qui peut avoir plusieurs valeurs $\{1, 2, ..., r\}$ est dite Multinomiale. L'apprentissage des paramètres de sa fonction de probabilité à partir de données est similaire au cas de variables de Bernoulli. L'estimateur de maximum de vraisemblance est donné par [161] :

$$p(X = k) = \frac{N_k}{N} \tag{3.27}$$

où N_k et le nombre d'occurrences de X_k dans $\mathcal{D}$.

La loi a priori conjuguées dans ce cas est une distribution de Dirichlet de paramètres $\{\alpha_1, ...\alpha_r\}$

Définition 1 (Distribution de Dirichlet [161])
Une distribution de Dirichlet d'ordre $K \geq 2$ avec les paramètres $\alpha_1, ..., \alpha_k > 0$, est définie par :

$$f(\theta_1, ..., \theta_{K-1}; \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} (\theta_i^{\alpha_i - 1}) \tag{3.28}$$

pour tout $\alpha_1, ..., \alpha_{K-1} > 0$ vérifiant $\theta_1 + ... + \theta_{K-1} < 1$, où θ_K est une abréviation pour $1 - \theta_1 - ... - \theta_{K-1}$ et $s = \sum \alpha_K$. Avec $B(\alpha)$ est une fonction Beta Multinomiale.

L'estimateur MAP de la fonction de probabilité de X est donné par :

$$p(X = k|D) = \frac{N_k + \alpha_k - 1}{\sum_k (N_k + \alpha_k - 1)} \tag{3.29}$$

3.6.3 Loi Multinoulli [79]

La loi Multinoulli ou Catégorique est une distribution sur une seule variable discrète avec k états différents, où k est fini. La distribution Multinoulli est paramétrée par un vecteur $\theta \in [0, 1]^k$, où θ_i donne la probabilité du $i - me$ état. Le dernier $k - me$ état de probabilité est donné par $1 - 1^T p$. On note que nous devons limiter $1^T \theta \leq 1$. Les distributions Multinoulli sont souvent utilisées pour se référer aux distributions sur les catégories des objets ou des scènes. La distribution Multinoulli est un terme qui a été introduit

récemment par Gustavo Lacerdo et popularisé par Murphy. La distribution de Multinoulli est un cas particulier de la distribution Multinomiale. Une distribution Multinomiale est la distribution sur les vecteurs dans $\{0, ..., n\}^k$ représentant le nombre de fois que chacune des k catégories est visitée lors de n échantillons sont tirés d'une distribution de Multinoulli. De nombreux textes utilisent le terme Multinomial pour désigner les distributions de Multinoulli sans clarifier qu'ils ne se réfèrent qu'au cas $n = 1$.

Les distributions Bernoulli et Multinoulli suffisent à décrire toute distribution sur leur domaine. Ils sont en mesure de décrire toute distribution au cours de leur domaine, pas tellement parce qu'ils sont particulièrement puissants, mais plutôt parce que leur domaine est simple ; ils modélisent des variables discrètes pour lesquelles il est possible d'énumérer tous les états. En ce qui concerne les variables continues, il existe beaucoup d'états, de sorte que toute distribution décrite par un petit nombre de les paramètres doivent imposer des limites strictes à la distribution.

3.7 Conclusion

Nous avons présenté dans ce chapitre les réseaux bayésiens. La construction d'un réseau bayésien est constituée de deux phases : une phase d'élaboration du modèle et une phase d'apprentissage de paramètres. Nous avons introduit les notions de l'inférence bayésienne qui permettent d'utiliser le modèle pour calculer des probabilités *a posteriori* à partir d'observations. Nous avons présenté également les distributions probabilistes les plus usuelles.

Nous détaillerons dans le chapitre suivant notre contribution pour l'annotation automatique des photos personnelles par les événements dans lesquels elles sont prises. Plus précisément, nous présentons notre modèle probabiliste pour la reconnaissance d'événements dans les albums de photos personnelles.

Chapitre 4
MODÉLISATION

4.1 Introduction

Ce chapitre est consacré à la modélisation et la construction de notre réseau bayésien pour la reconnaissance d'événements dans les albums photos. Cette contribution s'inscrit dans le contexte général de l'annotation sémantique d'images. L'objectif de notre modèle est d'annoter les albums avec des catégories d'événements en se basant sur l'analyse de leurs images en terme d'objets et de scènes. Dans notre travail, les images constituant un album sont représentées par des caractéristiques sémantiques correspondant aux objets et aux scènes. Nous commençons donc ce chapitre par la présentation de la phase d'extraction des caractéristiques d'images. Ensuite, nous détaillons notre modèle, l'apprentissage de ses paramètres et la catégorisation d'un nouvel album en utilisant l'inférence bayésienne avant de conclure.

4.2 Extraction des caractéristiques sémantiques d'une image

Avant de présenter en détail notre modèle graphique probabiliste pour la reconnaissance d'événements, dans les albums photos, nous commençons par présenter la méthode adoptée pour la représentation d'images. La première étape de notre approche consiste à extraire les caractéristiques d'images. Ces caractéristiques correspondent aux objets et scènes constituant l'image. Elles permettent de représenter l'image au niveau sémantique. Plusieurs méthodes ont été proposées pour la reconnaissance d'objets et la reconnaissance de scènes dans la littérature. Récemment, des excellentes performances ont été obtenues par les approches basées sur les réseaux de neurones convolutifs (CNNs). Les CNNs ont apporté une amélioration considérable aux performances des systèmes de reconnaissance d'objets et de reconnaissance de scènes ces dernières années. Cette conclusion est basée sur les résultats obtenus lors des challenges sur les bases de benchmark ImageNet [163] (pour la reconnaissance d'objets) et Places205 (pour la reconnaissance de scènes)

[104]. Nous avons alors adopté l'architecture du réseau *GoogleNet*. Ce dernier a eu les meilleurs résultats lors du défi ILSVRC 2014, pour l'extraction des caractéristiques d'images. Les techniques du transfer learning [110] permettent d'adapter l'architecture de GoogleNet, pré-entrainés sur les bases ImageNet et Places, pour l'extraction des objets et des scènes à partir des images. L'intérêt d'utiliser un modèle pré-entraîné est double : on utilise une architecture optimisée avec soin par des spécialistes, et l'on profite des capacités d'extraction de caractéristiques apprises sur un jeu de donnée riche comme ceux d'ImageNet et de Places205 [110].

4.3 Structure du modèle et génération de l'album photo

Dans cette section, nous fournissons les détails sur la structure du modèle graphique proposé et le processus de génération d'albums. Nous présentons aussi le schéma probabiliste de pertinence des caractéristiques utilisées pour améliorer la discrimination entre les différentes catégories d'événements. La structure de notre modèle génératif, illustrée dans la figure 4.1, correspond à un réseau bayésien où les nœuds représentent les variables et les arcs représentent les liens de dépendances entre elles. La notion qui est à la base des réseaux bayésiens est celle de la modularité : un système complexe est construit en combinant des parties plus simples [164]. La théorie des probabilité scelle l'ensemble en permettant de combiner les parties. Cela permet de modéliser un concept complexe comme l'évènement en fonction de concepts plus simples tels que les objets et les scènes. Par ailleurs, la théorie des graphes fournit des outils intuitifs permettant la modélisation par l'humain de systèmes complexes et la création d'algorithmes efficaces de raisonnement et d'apprentissage. De plus, Les réseaux bayésiens sont convenables pour représenter les relations causales et les dépendances entre les variables et leur combinaison. Ils permettent d'utiliser des données en provenance de multiples sources. Ces données servent à mettre à jour la connaissance et les croyances que l'on a sur le problème, de façon bayésienne, à partir d'un unique formalisme d'inférence et d'avoir ainsi une nouvelle vue du problème sachant les nouvelles données [164]. Ils sont également efficaces pour gérer les incertitudes provenant de la connaissance incomplète des dépendances dans les situations réelles [165].

La figure 4.1 illustre la représentation graphique de notre modèle bayé-

sien (PGM) pour la reconnaissance d'événements dans les albums photos. Les nœuds gris sont des variables observées dans les phases d'apprentissage et de test. Les nœuds blancs ne sont pas observés dans la phase de test. A l'extérieur du cadre *Album*, les nœuds avec traits discontinus sont des variables estimées. Enfin, les nœuds verts et les carrés verts avec coins arrondies sont les paramètres et les hyperparamètres du modèle, respectivement.

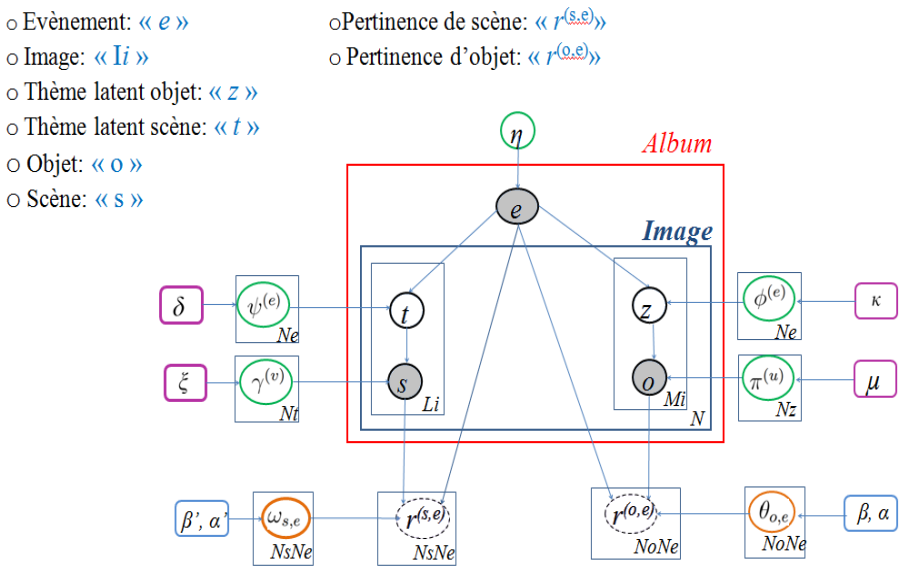

FIGURE 4.1 – Représentation graphique du modèle proposé. Les boîtes indiquent la réplication des variables aléatoires correspondantes : il y a N images dans un album $\mathcal{A}$, avec M_i objets observés et L_i scènes observées dans l'image I_i. Les variables e, o et s représentent, respectivement, les classes d'événement, d'objet et de scène. Les variables latentes pour générer l'objet et la scène sont respectivement z et t.

4.3.1 Processus génératif d'un album

Pour générer un album d'un événement particulier, nous commençons par choisir l'étiquette de la catégorie d'événement, par exemple un événement "*voyage*". Dans notre cas, nous supposons qu'un album est un ensemble d'images (non ordonnées) indépendantes les unes des autres étant donné la catégorie d'événement. De plus, pour faciliter notre modélisation, nous supposons que les objets et les scènes sont indépendants dans chaque image en fonction de la catégorie d'événement. Nous supposons également que dans une

catégorie d'événement, plusieurs *thèmes objet latents et thèmes scène latents* peuvent se manifester. Ces derniers représentent des contextes sémantiques dans lesquels les objets et les scènes apparaissent, respectivement, dans l'événement. Par exemple, dans un événement "*voyage*", nous pouvons considérer les thèmes objet "*voyage par avion*" et "*voyage par bateau*", et "*voyage sur littoral*" et "*voyage en montagne*" comme des thèmes scène. La sélection du thème objet "voyage par avion" privilégiera les objets qui se produisent fréquemment dans ce thème (par exemple, passeport, sac à dos, valise, etc.). De même, la sélection du thème scène "voyage sur littoral" favorisera les scènes fréquentes dans ce thème (par exemple plage, mer, etc.) comme montre la figure 4.2.

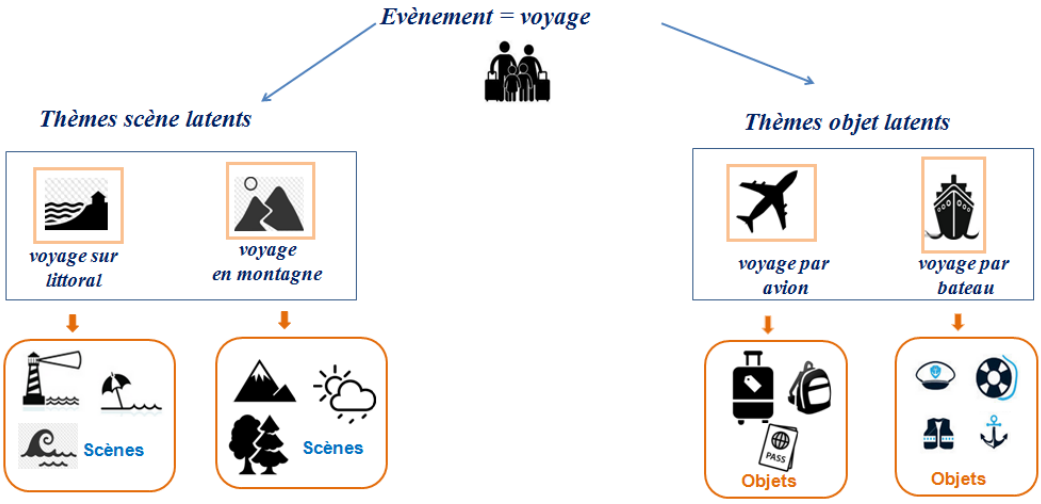

FIGURE 4.2 – Exemples de thèmes scène latents et thèmes objet latents présents dans l'évènement "voyage".

Pour générer un album d'images pour une catégorie d'événement particulière, nous générons d'abord des thèmes scène et thèmes objet à partir de mélanges de thèmes disponibles. Ensuite, en fonction des thèmes générés, nous générons les instances d'objets et de scènes constituant l'image. Plus formellement, soit un album $\mathcal{A}$ d'un événement particulier contenant un ensemble de N images dénotées par $\mathcal{A} = \{I_1, .., I_N\}$, où I_i est la i-ème image de l'album $\mathcal{A}$. Pour mieux comprendre notre modèle, nous passons par le processus génératif d'un album étant donné une catégorie d'événement particulière.

Nous commençons par définir un ensemble de notations que nous utiliserons dans la suite de chapitre. Soit N_e le nombre de catégories d'événements et soit N_o et N_s le nombre de différentes classes d'objets et de scènes, respectivement. De plus, soit $e \in \{1, 2, ..., N_e\}$ une variable aléatoire discrète représentant la catégorie d'événement. Une catégorie e est générée en fonction de la distribution $e \sim p(e|\eta)$, où η est un vecteur N_e-dimensionnel de paramètre d'une distribution Multinoulli. Nous supposons qu'un album appartient à une seule catégorie d'événement, et donc, nous générons un seul événement e. Pour chaque image $I_i \in \mathcal{A}$, les instances de scène et d'objet y appartenant peuvent être générées comme suit :

4.3.1.1 Processus génératif des scènes

Supposons qu'une instance de scène suit une distribution Multinoulli et qu'elle soit représentée par un vecteur N_s-dimensionnel, appelé s, avec les composantes $s^{(k)} \in \{0, 1\}$, où une seule composante est égal à un, avec $k \in \{1, 2, ..., N_s\}$. L'image I_i peut contenir L_i instances de scène $L_i \in \{1, ..., L_{max}\}$, qui sont représentées à l'aide d'un ensemble de vecteurs aléatoires $\mathbf{s}_i = \{s_{il}\}_{l=1,...,L_i}$, comme l'illustre la figure 4.3. Pour générer une instance de scène s_{il} dans la i-ème image d'album :

- On choisit un thème scène t_{il}, où t_{il} suit une distribution Multinoulli $\text{Mut}(\psi^{(e)})$ avec un vecteur N_t-dimensionnel de paramètres $\psi^{(e)}$. N_t est le nombre de thèmes dans l'espace latent de scène, et $t_{il}^{(v)} = 1$ indique que la v-ème composante est sélectionnée. Le vecteur $\psi^{(e)}$ a une distribution a priori de Dirichlet avec l'hyper-paramètre ξ.

- Une fois que la scène $v \in \{1, ..., N_t\}$ est sélectionnée, une instance de scène s_{il} est générée selon une distribution Multinoulli $\text{Mut}(\gamma^{(v)})$ avec un vecteur N_s-dimensionnel de paramètre $\gamma^{(v)}$. La variable s_{il} est un vecteur N_s-dimensionnel, où $s_{il}^{(k)}$ indique que l'instance de scène appartient à la k-ème classe. Enfin, le paramètre $\gamma^{(v)}$ a une distribution a priori de Dirichlet avec l'hyper-paramètre δ.

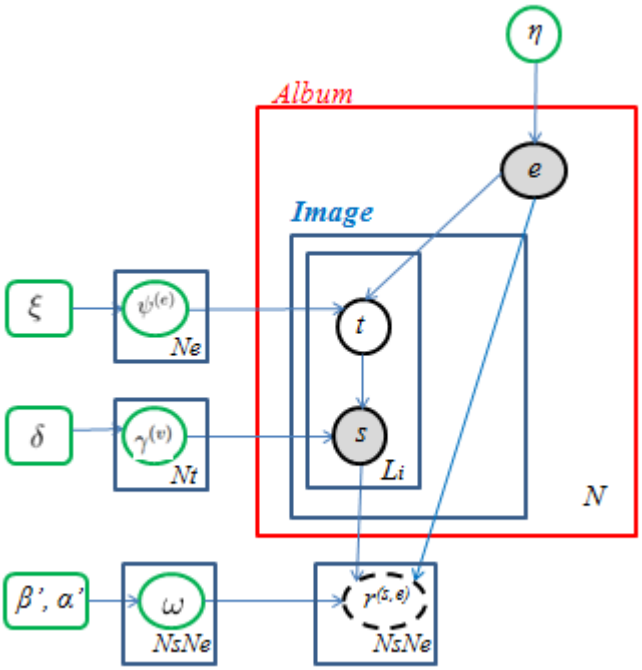

FIGURE 4.3 – Représentation graphique du modèle proposé basée uniquement sur les scènes.

4.3.1.2 Processus génératif des objets

Supposons qu'une instance d'objet suit une distribution Multinoulli et qu'elle est représentée par un vecteur N_o-dimensionnel, appelé o, avec les composantes $o^{(q)} \in \{0, 1\}$, où un seul composant est égal à un, avec $q \in \{1, 2, ..., N_o\}$. L'image I_i peut contenir M_i instances d'objet $M_i \in \{1, ..., M_{max}\}$, qui sont représentées à l'aide d'un ensemble de vecteurs aléatoires $\mathbf{o}_i = \{o_{im}\}_{m=1,...,M_i}$, comme l'illustre la figure 4.4. Pour générer une instance d'objet o_{im} dans la i-ème image d'un album :

- On choisit un thème objet z_{im}, où z_{im} suit une distribution Multinoulli Mut($\phi^{(e)}$) avec un vecteur N_z-dimensionnel $\phi^{(e)}$. N_z est le nombre de thèmes dans l'espace latent d'objet, et $z_{im}^{(u)} = 1$ indique que le u-ème thème est sélectionné. Le vecteur $\phi^{(e)}$ a une distribution a priori de Dirichlet avec l'hyper-paramètre κ.

- Une fois le thème objet $u \in \{1, ..., N_z\}$ est sélectionné, une instance d'objet o_{im} est générée selon une distribution Multinoulli Mut($\pi^{(u)}$) avec un

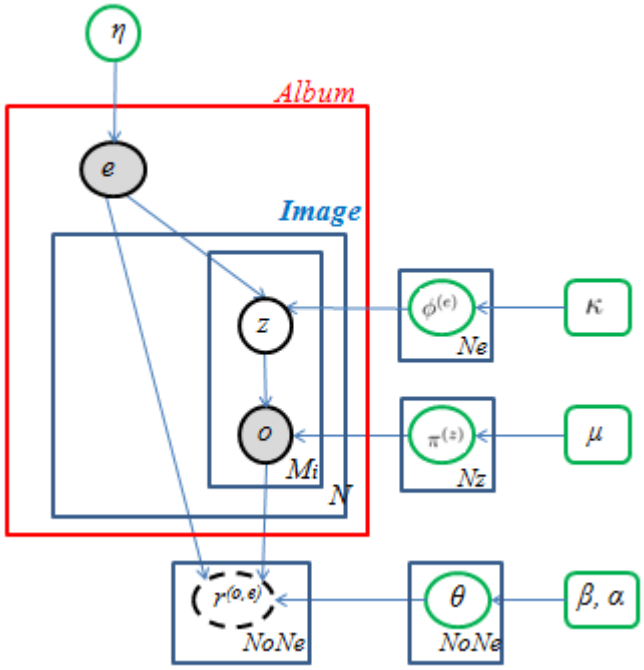

FIGURE 4.4 – Représentation graphique du modèle proposé basée uniquement sur les objets.

vecteur N_s-dimensionnel de paramètre $\pi^{(u)}$. La variable o_{im} est un vecteur N_o-dimensionnel, où $o_{im}^{(q)}$ indique que l'instance d'objet appartient à la q-ème classe d'objet. Enfin, les paramètres $\pi^{(u)}$ ont une distribution a priori de Dirichlet avec l'hyperparamètre μ.

4.3.1.3 Estimation de la pertinence objet/scène

Afin de prendre en considération la pertinence objet/scène pour la reconnaissance d'événements, nous introduisons de nouvelles variables $r^{(o,e)}$ et $r^{(s,e)}$ pour encoder la puissance de discrimination de chaque objet o et scène s par rapport à l'événement e, respectivement. Une fois que toutes les scènes et tous les objets ont été générés dans tous les événements considérés, nous pouvons estimer les paramètres de pertinence $r^{(o,e)}$ et $r^{(s,e)}$ comme suit :

1. Soit $r^{(o,e)}$ la variable aléatoire discrète binaire, où $r^{(o,e)} = 1$ si l'objet o est pertinent pour la catégorie d'événements e et $r^{(o,e)} = 0$, sinon. Alors, la probabilité $p(r^{(o,e)} = 1|e, o) = \theta_{o,e}$ (i.e., l'objet o est pertinent

pour l'événement e) est une distribution de Bernoulli Ber($\theta_{o,e}$) avec le paramètre $\theta_{o,e}$. Le paramètre $\theta_{o,e}$ a une distribution a priori Beta avec les hyperparamètres α et β.

2. Soit $r^{(s,e)}$ une variable aléatoire discrète binaire, où $r^{(s,e)} = 1$ si la scène s est pertinente pour la catégorie d'événements e et $r^{(s,e)} = 0$, sinon. Alors, la probabilité $p(r^{(s,e)} = 1|e, s) = \omega_{s,e}$ (i.e., la scène s est pertinente pour l'événement e) est une distribution de Bernoulli Ber($\omega_{s,e}$) avec le paramètre $\omega_{s,e}$. Le paramètre $\omega_{s,e}$ a une distribution a priori Beta avec les hyperparamètres α' et β'.

4.3.2 Probabilité jointe du modèle

Un modèle graphique probabiliste fournit une représentation compacte de la distribution de probabilité jointe d'un ensemble de variables. Pour trouver la probabilité jointe, nous considérons le processus génératif d'un album photo et l'ensemble des variables du modèle $\mathcal{X} = \{e, \mathbf{z}, \mathbf{t}, \mathbf{o}, \mathbf{s}, \mathbf{r_o}, \mathbf{r_s}\}$, avec l'événement $e \in \{1, ..., N_e\}$, les objets $\mathbf{o} = \{\mathbf{o}_{i=1:N}\}$, les scènes $\mathbf{s} = \{\mathbf{s}_{i=1:N}\}$, les thèmes objet $\mathbf{z} = \{\mathbf{z}_{i=1:N}\}$,les thèmes scène $\mathbf{t} = \{\mathbf{t}_{i=1:N}\}$, les variables de pertinence des différents objets $\mathbf{r_o} = \{r_{o,e|o=1:N_o,e=1:N_e}\}$, les variables de pertinences des différentes scènes $\mathbf{r_s} = \{r_{s,e|s=1:N_s,e=1:N_e}\}$.

La probabilité jointe de l'album sachant l'événement e peut être exprimée comme suit :

$$p(\mathcal{X}|\eta, \{\phi^{(e)}, \psi^{(e)}, \pi^{(u)}, \gamma^{(v)}, \theta_{o,e}, \omega_{s,e}\}) = p(e|\eta) \prod_{i=1}^{N} L_{i,o} \times L_{i,s} \tag{4.1}$$

où $L_{i,o}$ et $L_{i,s}$ sont les vraisemblances associés aux objets et aux scènes de la i-ème image. Elles sont définies par :

$$L_{i,o} = \prod_{m=1}^{M_i} p(z_{im}|\phi^{(e)}, e)p(o_{im}|z_{im}, \pi^{(u)})p(r^{(o,e)}|o_{im}, \theta_{o,e}, e) \tag{4.2}$$

$$L_{i,s} = \prod_{l=1}^{L_i} p(t_{il}|\psi^{(e)}, e)p(s_{il}|t_{il}, \gamma^{(v)})p(r^{(s,e)}|s_{il}, \omega_{s,e}, e) \tag{4.3}$$

4.4 Pertinence probabiliste des caractéristiques

Dans cette section, nous définissons les critères pour considérer une caractéristique, qu'il s'agisse d'une scène ou d'un objet, comme pertinente pour la discrimination des catégories d'événements. En d'autres termes, nous passons par le processus d'apprentissage pour estimer les paramètres $\{\theta_{o,e}, \omega_{s,e}\}$, respectivement. Une caractéristique est pertinente pour une catégorie d'événement si elle contribue à sa discrimination.

La pertinence des caractéristiques peut être déterminée par une mesure de dépendance entre une variable aléatoire et une classe cible. Parmi ces mesures, l'*information mutuelle* (MI) quantifie les informations incorporées dans une caractéristique d'entrée donnée pour prédire une classe cible [166].

En fait, un objet/scène pertinent peut réduire l'incertitude et apporter des connaissances sur une catégorie d'événement. Par exemple, la présence des objets "arbre de Noël" ou "océan", respectivement, est plus informatif sur les événements "*Noël*" et "*Croisière*" que d'autres objets.

Plus formellement, pour chaque catégorie d'événements $e \in \{1, ..., N_e\}$, nous définissons une variable binaire $r^{(x,e)}$ avec $x \in \{o, s\}$ et l'objet $o \in \{1, ..., N_o\}$ et la scène $s \in \{1, ..., N_s\}$. La variable $r^{(x,e)}$ prend la valeur 1 si la caractéristique x est pertinente pour la catégorie d'événements e et prend la valeur 0 dans le cas contraire. Par conséquent, la variable discrète $r^{(x,e)}$ suit une distribution de Bernoulli. Dans notre cas, étant donné un échantillon de plusieurs albums tirés de plusieurs événements, nous pouvons calculer l'information mutuelle (MI) pour chaque objet/scène par rapport à chaque catégorie d'événement, en utilisant la formule suivante [167] :

$$MI(x, e) = \sum_{x \in \{0,1\}} \sum_{a=e, a \neq e} p(x, a) \log \left(\frac{p(x, a)}{p(x)p(a)} \right) \tag{4.4}$$

Lorsque l'information mutuelle entre la caractéristique d'entrée x et la classe cible e est faible, la caractéristique d'entrée n'est pas pertinente pour la classe cible, quel que soit l'algorithme de classification. Par contre, des valeurs plus élevées de l'information mutuelle signifient que la caractéristique d'entrée est pertinente pour la classe cible.

Notons que pour catégoriser les événements, les gens comptent souvent sur des objets représentatifs/discriminatifs (par exemple "*toge*" pour identifier

l'événement "*fête d'obtention de diplôme*", etc.), tandis que l'absence d'objets n'est pas très informative sur l'événement(par exemple l'absence de "*bateau*" ne signifie pas nécessairement un événement de *fête d'obtention de diplôme*). Pour inclure cet aspect dans la détermination de la pertinence des caractéristiques pour la reconnaissance d'événements, nous utilisons l'analyse de corrélation entre les caractéristiques (objets/scènes) et les catégories d'événements. Le coefficient de corrélation $\rho(x, e) \in [-1, +1]$ mesure la force de cette relation. $\rho(x, e) > 0$ (resp. $\rho(x, e) < 0$) indique une relation de corrélation positive (resp. négative) entre un objet/scène et une catégorie d'événement e. Par conséquent, de grandes valeurs de $\rho(x, e)$ et MI indiquent une forte pertinence de la caractéristique x pour la prédiction de la catégorie d'événement e. Nous proposons la fonction suivante pour approximer l'observation de la variable $r^{(x,e)}$ dans un échantillon donné d'albums :

$$r^{(x,e)} \approx 1 - \exp\left[-MI(x, e)\, H(\rho(x, e))\right], \tag{4.5}$$

où $H(.)$ est la fonction Heaviside qui considère uniquement les cas de valeurs positives de corrélation afin d'attribuer des valeurs de pertinence plus élevées. En utilisant la formule ci-dessus, les valeurs positives de corrélation et les valeurs élevées de l'information mutuelle donnent $r^{(x,e)} \to 1$. Pour les valeurs négatives de corrélation et/ou les valeurs faibles de l'information mutuelle, $r^{(x,e)} \to 0$.

4.5 Estimation des paramètres du modèle

Dans cette section, nous présentons la procédure d'estimation des paramètres du modèle proposé. En premier lieu, nous introduisons l'apprentissage des paramètre d'événement, d'objet et de scène. En second lieu, nous présentons la méthode d'estimation des paramètres de pertinence de scènes et d'objets.

4.5.1 Estimation des paramètres d'événement, d'objet et de scène

Dans ce qui suit, nous décrivons la stratégie d'apprentissage des paramètres $\{\pi^{(u)}, \phi^{(e)}, \psi^{(e)}, \gamma^{(v)}\}$ de notre modèle, tel qu'illustré en la figure 4.1, relatifs aux événements, thème objet, objets, thèmes scènes, scènes, respectivement.

Nous supposons que les événements sont équiprobables. Ils suivent une distribution a priori uniforme vu qu'il n y a pas de raison qu'un événement soit plus important qu'un autre, avec $p(e) = 1/N_e$. Nous supposons également que les objets et les scènes sont indépendants sachant l'événement e. Par conséquent, les paramètres d'objet $\{\phi^{(e)}, \pi^{(u)}\}$ et les paramètres de scène $\{\psi^{(e)}, \gamma^{(v)}\}$ peuvent être appris séparément.

Nous commençons par les caractéristiques objet, la distibution des thèmes z_{im} sachant une catégorie d'événements e est représentée par le vecteur N_z-dimensionnel $\phi^{(e)}$. Elle suit une distribution a priori de Dirichlet avec l'hyper-paramètre κ. L'estimation bayésienne des entrées du vecteur $\phi^{(e)}$ est donnée comme suit

$$\phi_u^{(e)} = p(z_{im}^{(u)} = 1|e) = \frac{n_{u,e} + \kappa_u}{\sum_{j=1}^{N_z} n_{j,e} + N_z \times \kappa_u} \tag{4.6}$$

où $n_{u,e}$ est le nombre d'occurrences du thème objet $z_{im}^{(u)}$ dans la catégorie événement e ; et κ_u est la u-ème entrée du vecteur κ.

La distribution de l'objet o étant donné le thème $z_{im}^{(u)} = 1$, est représentée par $\pi^{(u)}$. Ce dernier vecteur $\pi^{(u)}$ suit une distribution a priori de Dirichlet avec l'hyper-paramètre μ. L'estimation bayésienne pour $\pi^{(u)}$ est donnée comme suit :

$$\pi_q^{(u)} = p(o_{im}^{(q)} = 1|z_{im}^{(u)} = 1) = \frac{n'_{q,u} + \mu_u}{\sum_{j=1}^{N_o} n'_{j,u} + N_o \times \mu_u}, \tag{4.7}$$

où $n'_{q,u}$ est le nombre d'occurrences de la classe objet $o_{im}^{(q)}$ sachant le thème $z_{im}^{(u)}$, et μ_u est la u-ème entrée du vecteur μ.

De même pour la caractéristique de scène, nous avons le vecteur N_z-dimensionnel $\psi^{(e)}$ représentant la distribution du thème scène t sachant l'événement e. $\psi^{(e)}$ suit une distribution a priori de Dirichlet avec l'hyperparamètre ξ ; et peut être calculé par :

$$\psi_v^{(e)} = p(t_{il}^{(v)} = 1|e) = \frac{m_{v,e} + \xi_v}{\sum_{j=1}^{N_t} m_{j,e} + N_t \times \xi_v} \tag{4.8}$$

où $m_{v,e}$ est le nombre d'occurrences du thème scène $t_{il}^{(v)}$ étant donné l'événement e et ξ_u est la v-ème entrée du vecteur ξ.

La distribution de scène s étant donné le thème $t_{il}^{(v)} = 1$ est représentée par $\gamma^{(v)}$. Nous supposons que $\gamma^{(v)}$ suit une distribution a priori de Dirichlet

avec l'hyperparamètre δ. L'estimation bayésienne pour $\gamma^{(v)}$ est donnée comme suit :

$$\gamma_k^{(v)} = p(s_{il}^{(k)} = 1 | t_{il}^{(v)} = 1) = \frac{m'_{k,v} + \delta_v}{\sum_{j=1}^{N_s} m'_{j,v} + N_s \times \delta_v} \tag{4.9}$$

où $m'_{k,v}$ est le nombre d'occurrences de la scène $s_{il}^{(k)}$ lorsque le thème $t_{il}^{(v)}$ est sélectionné et δ_v est la v-ème entrée du vecteur δ. Enfin, selon [168], le choix intuitif pour les entrées hyperparamètres $\{\kappa, \mu, \xi, \delta\}$ est la distribution a priori uniforme.

4.5.2 Estimation des paramètres de pertinence

Le but de cette phase est d'estimer les paramètres $\theta_{o,e}$ et $\omega_{s,e}$ des distributions de Bernoulli $Ber(\theta_{o,e})$ et $Ber(\omega_{s,e})$ qui expriment les pertinences d'objet o et de scène s étant donné la catégorie d'événement e. Sans perte de généralité, nous développons la formulation pour l'estimation de pertinence d'objet o. Les mêmes étapes peuvent être suivies pour estimer la pertinence de scène s. La distribution de Bernoulli pour la variable $r^{(o,e)}$ est formulée comme suit :

$$p(r^{(o,e)}) = \theta_{o,e}^{r^{(o,e)}} (1 - \theta_{o,e})^{(1 - r^{(o,e)})} \tag{4.10}$$

Ensuite, nous supposons que nous avons tiré T échantillons de multiples ensembles de données, où chaque échantillon contient plusieurs albums d'événements différents. L'objectif est d'estimer, à partir de plusieurs échantillons, la probabilité de pertinence de chaque objet et scène par rapport à une catégorie d'événement. En d'autres termes, pour estimer $\theta_{o,e}$, nous faisons plusieurs tirages indépendants de la variable $r^{(o,e)}$ selon l'équation (4.5) et nous utilisons l'estimation bayésienne sur la base de sa distribution *a priori* Beta. Plus formellement, soit $\mathcal{D} = \{r_1^{(o,e)}, \ldots, r_T^{(o,e)}\}$ un ensemble de T observations indépendantes et identiquement distribuées *iid* de la variable $r^{(o,e)}$. On peut facilement montrer que l'estimation du maximum de vraisemblance pour $\theta_{o,e}$ est donné par :

$$\hat{\theta}_{o,e} = \arg \max_{\{\theta_{o,e}\}} p(\mathcal{D} | \theta_{o,e}) \tag{4.11}$$

où $p(\mathcal{D}|\theta_{o,e})$ est la vraisemblance des paramètres donnée par la factorisation suivante :

$$\begin{aligned} p(\mathcal{D}|\theta_{o,e}) &= \prod_{i=1}^{T} p(\{r_i^{(o,e)}|\theta_{o,e}) \\ &= \prod_{i=1}^{T} \theta_{o,e}^{r_i^{(o,e)}} (1-\theta_{o,e})^{1-r_i^{(o,e)}} \\ &= \theta_{o,e}^{N_1}(1-\theta_{o,e})^{N_2} \end{aligned} \tag{4.12}$$

où $N_1 = \sum_i^T r_i^{(o,e)}$ et $N_2 = (T - N1)$ donne le nombre d'échantillons où l'objet est jugé pertinent pour la catégorie d'événement e.

Pour considérer une estimation bayésienne pour le paramètre $\theta_{o,e}$, nous supposons une distribution a priori Beta pour $\theta_{o,e}$ avec les hyperparamètres α et β. Cette distribution a priori conjuguée est considérée pour faciliter le calcul. L'estimation MAP pour $\theta_{o,e}$ est alors calculée comme suit :

$$\begin{aligned} \hat{\theta}_{o,e} &= \arg\max_{\{\theta_{o,e}\}} \left[Beta(\alpha + N_1 - 1, \beta + N_2 - 1)\right] \\ &= \frac{\alpha + N_1 - 1}{\alpha + \beta + T - 2} \end{aligned} \tag{4.13}$$

4.6 Prédiction d'événements pour les nouveaux albums

La reconnaissance d'un événement à partir d'un album photo est considérée comme un problème d'inférence dans les réseaux bayésiens. Le but est d'estimer les valeurs des nœuds cibles (non observés) à partir des valeurs des nœuds observés. Étant donné un nouvel album $\mathcal{A} = \{I_1, I_2, \ldots, I_N\}$, l'objectif est de trouver sa catégorie d'événement e. Pour cela, nous calculons la probabilité a posteriori maximale (MAP) en fonction de chaque catégorie d'événement $e \in \{1, 2, \ldots, N_e\}$, avec N_e est le nombre de catégories d'événements. Ensuite, nous attribuons l'étiquette $\hat{e}$ à l'album $\mathcal{A}$ selon l'équation suivante :

$$\hat{e} = \arg\max_{e} p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) \tag{4.14}$$

où $\Theta = (\eta, \{\phi^{(e)}, \psi^{(e)}, \pi^{(u)}, \gamma^{(v)}, \theta_{o,e}, \omega_{s,e}\})$ indique l'ensemble des paramètres du modèle relatifs aux événement, thème objet, thème scène, objet, scène, pertinence d'objet et pertinence de scène, respectivement. Sans perte de gé-

néralité, nous supposons que les images de l'album $\mathcal{A}$ sont indépendantes et que la probabilité d'une catégorie d'événement dépend uniquement des paramètres de cette catégorie. En écrivant l'album en termes d'images, l'équation. (4.14) devient :

$$\begin{aligned} p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) &= p(e|I_1, I_2, ..., I_N, \mathbf{r_s}, \mathbf{r_o}, \Theta) \\ &\propto \prod_{i=1}^{N} p(e|I_i, \mathbf{r_s}, \mathbf{r_o}, \Theta, e) \end{aligned} \tag{4.15}$$

Les unités de base d'une image I_i sont des scènes et des objets, où $I_i = \{\mathbf{o}_i, \mathbf{s}_i\}$. Ainsi, on peut étendre le terme droit de l'équation (4.15) et exprimer la probabilité d'une image en prenant compte de ses composantes (objet et scène) :

$$p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) = \prod_{i=1}^{N} p(e|\mathbf{o}_i, \mathbf{s}_i, \mathbf{r_s}, \mathbf{r_o}, \Theta) \tag{4.16}$$

Par ailleurs, nous supposons que les objets et les scènes de chaque image sont détectés séparément dans des phases indépendantes. Nous supposons également que les objets $\mathbf{o}_i = \{o_{im}, \text{où } m \in \{1, \ldots, M_i\}\}$ et les scènes $\mathbf{s}_i = \{s_{il}, \text{où } l \in \{1, \ldots, L_i\}\}$ d'une image I_i sont générés indépendamment les uns des autres. Enfin, nous supposons que r_s et r_o sont indépendants. Par conséquent, nous obtenons :

$$\begin{aligned} p(e|\mathcal{A}, \mathbf{r_s}, \mathbf{r_o}, \Theta) &= \prod_{i=1}^{N} p(e|\mathbf{o}_i, \mathbf{r_s}, \mathbf{r_o}, \Theta) p(e|\mathbf{s}_i, \mathbf{r_s}, \mathbf{r_o}, \Theta) \\ &= \prod_{i=1}^{N} \prod_{m=1}^{M_i} p(e|o_{im}, r^{(o,e)}, \Theta) \prod_{l=1}^{L_i} p(e|s_{il}, r^{(s,e)}, \Theta) \end{aligned} \tag{4.17}$$

Par la suite, le premier terme de l'équation (4.17), qui représente la vraisemblance d'un album au niveau des objets, peut être marginalisé sur toutes les

valeurs des thèmes objets, comme suit :

$$\begin{aligned} p(e|o_{im}, r^{(o,e)}, \Theta) &\propto p(o_{im}, r^{(o,e)}, \Theta|e)p(e|\eta) \\ &\propto p(r^{(o,e)}|o_{im}, e, \Theta)p(o_{im}|e, \Theta)p(e|\eta) \\ &\propto p(r^{(o,e)}|o_{im}, e, \Theta)\sum_{z_{im}} p(o_{im}|z_{im}, e, \Theta)p(z_{im}|e, \Theta)p(e|\eta) \end{aligned} \tag{4.18}$$

L'expansion des termes individuels de l'équation (4.18) dépend de la structure de notre modèle illustrée dans la figure 4.1. Le terme $p(r^{(o,e)}|o_{im}, e, \Theta)$ est la probabilité que l'objet o_{im} soit pertinent pour la classe d'événements e. La probabilité de co-occurrence de l'objet o_{im} dans l'événement e est représentée par $p(o_{im}, e, \Theta) = p(o_{im}|e, \Theta)p(e|\eta)$, où $p(e|\eta)$ est la probabilité a priori de l'événement e. De même que le premier terme de l'équation (4.17), le second terme de la même équation peut être marginalisé sur toutes les valeurs de thèmes scène, comme suit :

$$\begin{aligned} p(e|s_{il}, r^{(s,e)}, \Theta) &\propto p(s_{il}, r^{(s,e)}, \Theta|e)p(e|\eta) \\ &\propto p(r^{(s,e)}|s_{il}, e, \Theta)p(s_{il}|e, \Theta)p(e|\eta) \\ &\propto p(r^{(s,e)}|s_{il}, e, \Theta)\sum_{t_{il}} p(s_{il}|t_{il}, e, \Theta)p(t_{il}|e, \Theta)p(e|\eta) \end{aligned} \tag{4.19}$$

où $p(r^{(s,e)}|s_{il}, e, \Theta)$ est la probabilité que la scène s_{il} soit pertinente pour la classe d'événements e. La probabilité de co-occurrence de la scène s_{il} dans l'événement e est représentée par $p(s_{il}, e, \Theta) = p(s_{il}|e, \Theta)p(e|\eta)$, où $p(e|\eta)$ est la probabilité a priori de l'événement e.

4.7 Conclusion

Dans ce chapitre, nous avons présenté notre réseau bayésien pour la reconnaissance d'événements dans les albums photos. L'album est représenté par des caractéristiques de haut niveau. Ces caractéristiques sont représentées par des objets et des scènes des images qui constituent l'album. Les objets et les scènes apportent des informations complémentaires sur l'album. Les objets représentent les acteurs principaux dans un événement, tandis que les scènes représentent le contexte dans lequel ces acteurs apparaissent. Nous avons uti-

lisé les réseaux de neurones convolutionnels pour l'extraction des objets et des scènes. Le réseau bayésien proposé combine ces caractéristiques pour pouvoir prédire l'événement dans lequel les photos ont été prises. Par ailleurs, nous avons introduit la notion de pertinence pour encoder le pouvoir discriminatif des caractéristiques afin d'augmenter la précision de prédiction d'événements. Nous avons détaillé la procédure d'estimation des paramètres du modèle. Nous avons montré l'utilisation de l'inférence bayésienne pour la classification d'un nouvel album dans une catégorie d'événement. Dans le chapitre suivant, la validation de notre modèle est présentée à travers un protocole expérimental.

Chapitre 5
EXPÉRIMENTATION ET ÉVALUATION

5.1 Introduction

Ce chapitre est consacré à la présentation des différents tests menés dans le cadre d'expérimentation afin de valider le modèle que nous avons proposé pour la reconnaissance d'événements dans les albums. Nous décrivons d'abord les jeux de données ainsi que les différentes mesures d'évaluation utilisées pour effectuer nos tests. Nous présentons la toolbox Caffe utilisée pour exploiter les réseau de neurones convolutionnels pour la reconnaissance d'objets et de scènes afin de représenter les images au niveau sémantique. Nous validons séparément les différents modules constituant notre modèle, à savoir les modules utilisant des objets et des scènes séparément pour la reconnaissance d'événements. Nous étudions l'impact de l'utilisation de la pertinence des caractéristiques sur les performances de la reconnaissance d'événements. Enfin, nous évaluons les performances de l'approche proposée en la comparant à d'autres approches de reconnaissance d'événements proposées dans la littérature, et nous discutons les résultats obtenus avant de conclure.

5.2 Présentation des bases de tests

Dans cette section, nous présentons les différentes bases d'images utilisées pour effectuer nos tests. La première base présentée est celle exploitée pour la validation de la reconnaissance d'événements dans les albums. Les deux autres bases d'images ont été utilisées pour l'extraction des caractéristiques sémantiques d'images.

5.2.1 Personal Event Collections (PEC) dataset

Afin de valider notre approche, nous avons utilisé la base de collections de photos personnelles "Personal Event Collections (PEC) dataset [1]". Elle a été

1. https://www.vision.ee.ethz.ch/datasets_extra/pec/

proposée comme benchmark en raison du manque de bases de références de photos personnelles permettant l'évaluation et la comparaison des différentes méthodes de reconnaissance d'événements. La base PEC a été crée par Bossard et al. [38] en 2013 dans le but de proposer une base référence dans les domaines d'annotation et de catégorisation des collections de photos personnelles. Cette base est constituée de 807 collections qui sont composées de 61364 images. Ces collections ont été collectées sur internet, plus précisément de Flicker, suivant des directives bien définies. Ces images correspondent à 14 catégories d'événements sociaux :

- Birthday (Anniversaire).
- Children's birthday (Anniversaire d'enfants).
- Christmas (Noël).
- Concert (Concert).
- Cruise (Croisière).
- Easter (Pâques).
- Exhibition (Exposition).
- Graduation (Fête de remise des diplômes).
- Halloween (Halloween).
- Hiking (Randonnée).
- Road trip (Tour en voiture).
- Saint Patrick's day (Jour de Saint Patrick).
- Skiing (Ski).
- Wedding (Mariage).

Chaque collection représente un album photos correspondant à un seul événement. L'annotation d'événements est définie au niveau de l'album. Les albums photos sont créés par des utilisateurs différents à travers le monde. Les photos sont prises dans des conditions réalistes avec des changements d'échelles,

Classe	Nombre d'albums	Nombre de photos
Birthday	60	3227
Children's Birthday	64	3714
Christmas	75	4118
Concert	43	2565
Boat cruise	45	4983
Easter	84	3962
Exhibition	70	3032
Graduation	51	2532
Halloween	40	2403
Hiking	49	2812
Road Trip	55	10469
St.Patrick's Day	55	5082
Skiing	44	2512
Wedding	69	9953
Total	**807**	**61364**

TABLE 5.1 – Nombre d'albums, d'images d'apprentissage et de test par classe constituant la base PEC [38].

d'illumination et différents points de vue. Le lien de téléchargement de la base est fourni sur demande en raison de la nature personnelle des photos, et pour respecter la confidentialité des propriétaires de photos. Le tableau 5.1 représente la distribution du nombres d'albums et d'images par classe d'événements.

La figure 5.1 présente quelques photos de cette base. Dans notre évaluation, nous avons utilisé le même protocole de validation suggéré par Bossard et al. [38] associé à la base PEC. Pour les tests, dix albums photos par classe sont utilisés (140 albums au total). Pour l'apprentissage des paramètres de notre modèle, nous avons sélectionné aléatoirement six albums par classe d'événement (84 albums au total) parmi l'ensemble d'apprentissage proposé. Dans le reste de ce chapitre, nous allons garder les labels des catégories d'événements originaux en anglais.

FIGURE 5.1 – Exemples d'images de la base PEC [38].

5.2.2 La base ImageNet

Pour l'apprentissage du réseau convolutionnel des objets avec la toolbox caffe, nous nous sommes intéressés à la base d'images ImageNet [163] qui a été crée et mise à la disposition des chercheurs dans le cadre d'organisation des compétitions internationales. Cette base permet aux chercheurs de confronter leurs modèles de classification dans des tâches de vision par ordinateur, notamment dans la reconnaissance d'objets et la localisation d'objets. Elle est à ce jour l'une des bases les plus répandus dans le domaine de la reconnaissance et localisation d'objets. Cette base est constituée de 1.2 million d'images réparties en images d'apprentissage et images de test représentant en tout 1000 classes d'objets. La figure 5.2[2] illustre des exemples d'images de cette base.

5.2.3 La base Places205

Pour l'apprentissage du réseau convolutionnel des scènes avec la toolbox caffe, nous nous sommes intéressés à la base d'images Places205. Cette base décrite dans [104], contient 2,5 millions images constituant 205 catégories de

2. Base ImageNet : http://www.image-net.org/challenges/LSVRC/2010/

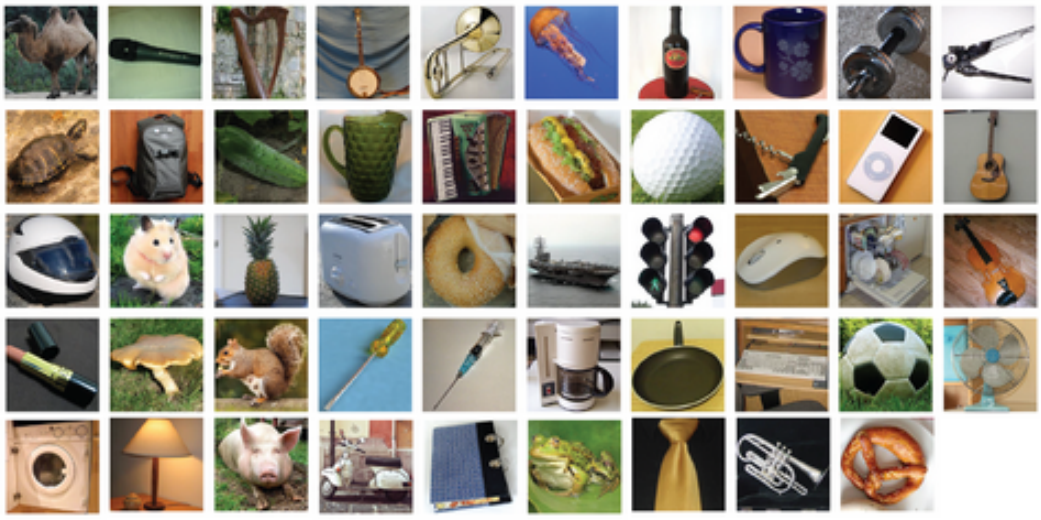

FIGURE 5.2 – Exemples d'images de la base ImageNet.

scènes. Le nombre d'images par classe varie entre 5.000 et 15.000. L'ensemble d'apprentissage contient 2,448,873 images au total, avec 100 images par catégorie pour l'ensemble de validation et 200 images par catégorie pour l'ensemble de test. La figure 5.3 [3] illustre des exemples de cette base.

5.3 Critères d'évaluation

Nous introduisons dans cette section les différentes mesures d'évaluation de performances utilisées lors de nos expérimentations :

- **Matrice de confusion [169]** : Une erreur souvent commise par le système de classification est la confusion entre les classes. Pour évaluer quantitativement ces erreurs, nous utilisons les matrices de confusion qui rassemblent les taux d'erreurs dues aux confusions pour chacune des classes.

- **Courbe de précision/rappel [170]** : La qualité de détection d'objet est évaluée dans notre travail par la courbe donnant la précision en fonction du rappel. Cette courbe permet de mettre en évidence deux aspects de détection :

3. Base Places205 : http://places.csail.mit.edu/browser.html

FIGURE 5.3 – Exemples de scènes de la base Places205.

  - le taux de précision ou le nombre d'objets pertinents retrouvés rapporté au nombre d'objets total détecté par l'algorithme pour une requête donnée.
  - le taux de rappel ou la proportion d'objets trouvés parmi tous les objets pertinents présents dans la base de données.

- **F score ou F mesure [171]** : Un système est caractérisé par une courbe ou par un couple (rappel/précision). On peut aussi ajouter un autre critère d'évaluation qui combine entre la précision et le rappel qui est le F score. Ce dernier est un indicateur de synthèse communément utilisé pour évaluer les algorithmes de classification, à partir des deux mesures précision et rappel. Elle est définie par :

$$F = 2\frac{precision \times rappel}{precision + rappel} \tag{5.1}$$

- **Précision moyenne [172]** : Afin de pouvoir statuer clairement si une méthode est meilleure qu'une autre et de combien, il est nécessaire de pouvoir présenter les performances d'une méthode par une valeur numérique unique. La mesure de précision moyenne (Average Precision-

AP) est mesurée généralement pour comparer les performances de deux méthodes sur un ensemble de classes. L'AP est une approximation de l'aire sous la courbe de précision/rappel. La moyenne arithmétique des APs est appelée la moyenne des précisions moyennes (Mean Average Precision-MAP).

5.4 Représentation d'image

Afin d'entraîner le réseau convolutionnel pour la reconnaissance d'objets, nous avons conçu une nouvelle base d'objets. Elle est constituée de 68% d'objets pris de la base ImageNet, tandis que les images des 32% des objets restants sont collectées sur internet. Cette nouvelle base contient les objets considérés comme discriminatifs pour les classes d'évènements de la base PEC. La toolbox Caffe [173] est utilisée pour la reconnaissance des instances d'objets et de scènes. Caffe est une toolbox pour l'apprentissage profond (en anglais Deep Learning). Elle est développée par l'équipe Berkeley AI Research (BAIR) de l'université de Berkeley. Cette toolbox permet l'apprentissage de nouveaux modèles de classification et l'extraction des caractéristiques de bas niveaux en exploitant des modèles de réseaux de neurones convolutionnels pré-entrainés par la technique de *fine tuning*. Motivés par les résultats prometteurs que Wang et al. [29] ont obtenu pour la reconnaissance d'événements culturels sur la base d'images 'Looking At People' à l'atelier CVPR, nous avons adopté l'architecture GoogLeNet pour construire nos réseaux convolutionnels objet et scène. Les détails de l'architecture du réseau GoogLeNet sont fournis dans l'article [109].

Le modèle GoogLeNet, initialisé sur la base des objets ImageNet, est utilisé pour entraîner le réseau convolutionnel objet. Pour cela, nous avons effectué un fine tuning sur notre nouvelle base d'objets. Durant la phase d'apprentissage, toutes les images ont été redimensionnées à 256×256 pixels. Pour la reconnaissance de scène, nous avons utilisé le modèle GoogLeNet pré-entraîné sur la base d'images de scène Places205. Dans la phase de test, pour les réseaux convolutionnels de scène et d'objet, nous avons redimensionné les images à 256×256 pixels avant de les faire passer, comme entrées, aux réseaux.

5.5 Évaluation des modules scène/objet de réseau bayésien

Dans cette section, nous évaluons individuellement la contribution des caractéristiques scènes/objets pour la reconnaissance d'événements dans les albums photos. Pour cela, nous avons implémenté deux versions de l'équation 4.17. La première version appelée (O+PGM) ne contient que des informations sur les objets et ignore les termes scènes dans l'équation 4.17. La deuxième version appelée (S+PGM) ne contient que des informations sur les scènes et ignore les termes objets de l'équation 4.17.

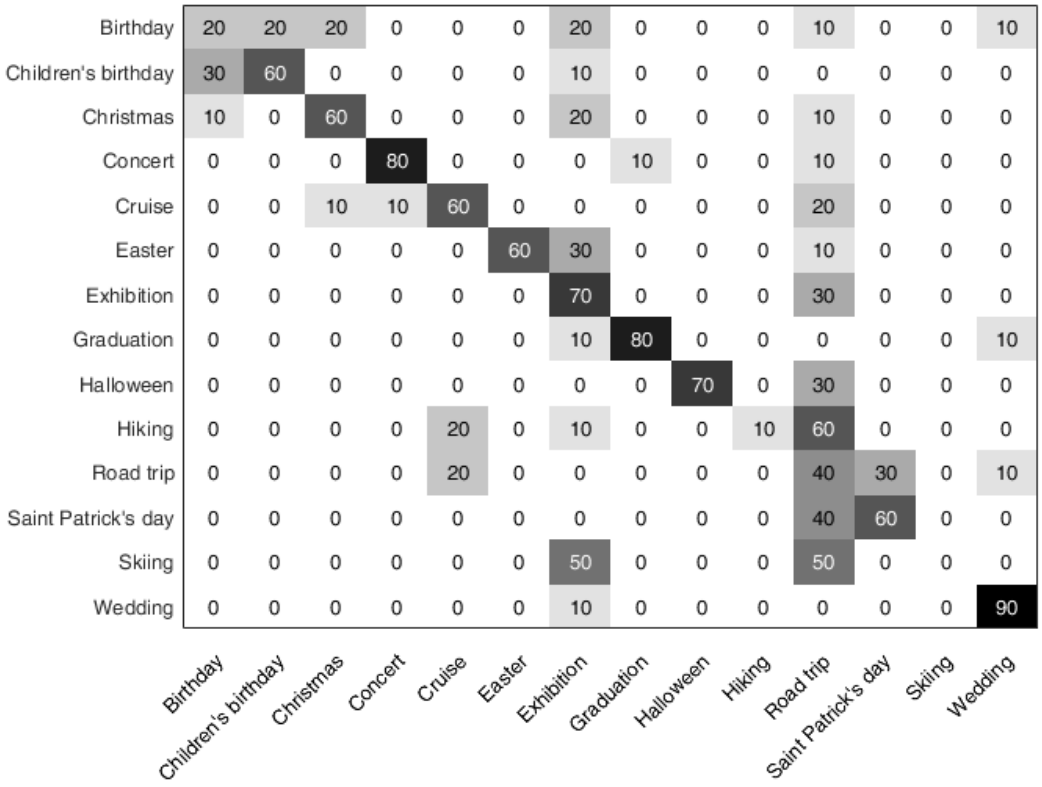

FIGURE 5.4 – Matrice de confusion pour reconnaissance d'événements basés uniquement sur les objets (O+PGM).

Nous évaluons la performance de la reconnaissance d'événements en utilisant des matrices de confusion. Les figures 5.4 et 5.5 présentent les matrices obtenues pour les deux versions de notre modèle (O+PGM) et (S+PGM), où les lignes correspondent aux prédictions retournées par le système et les colonnes correspondent aux étiquettes de la vérité terrain, respectivement.

Nous pouvons noter que le modèle basé sur les objets (O+PGM) produit de meilleurs résultats que celui basé sur les scènes (S+PGM) avec une précision de $54.28\%, 41.42\%$, respectivement.

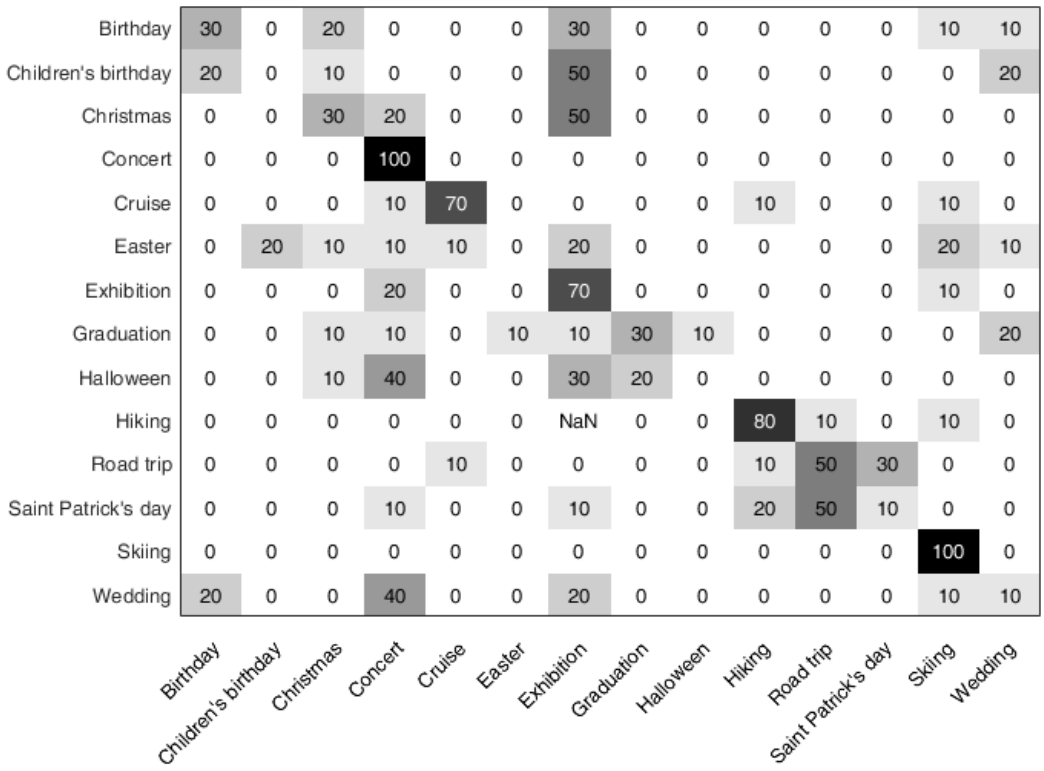

FIGURE 5.5 – Matrice de confusion pour la reconnaissance d'événements en utilisant uniquement les scènes pour la reconnaissance d'évènements (S+PGM).

En général, les objets et les scènes constituent de bonnes caractéristiques pour la reconnaissance des événements lorsque ces derniers peuvent être discriminés soit par des objets clés ou par des scènes clés. Par exemple, l'événement "Wedding" est facilement reconnaissable lorsque l'objet "wedding dress (robe de mariée)" se trouve dans l'album photo. Nous remarquons, cependant, que certaines paires de classes d'événements peuvent être facilement confondues, par exemple l'événement "Road Trip" et "Hiking", ce qui est principalement dû à la forte similitude visuelle de ces événements.

En outre, certaines erreurs de classification d'événements sont dues aux erreurs de détection des caractéristiques clés. Afin de mieux étudier ce problème,

nous avons analysé le taux d'erreurs de détection d'objet "boat (bateau)" dans un échantillon de 403 images, tiré à partir des albums de l'événement "Cruise" de la base PEC. La figure 5.6 montre la courbe de précision/rappel pour la détection d'objet "bateau". Les meilleurs valeurs de précision, rappel et F score sont 0.72, 0.78 et 0.75, respectivement. Nous avons remarqué que les prédictions erronées sont dues principalement à l'utilisation de l'architecture GoogLeNet pour la reconnaissance d'objets. L'utilisation de meilleurs modèles de CNNs, ou la combinaison de plusieurs modèles CNNs comme indiqué dans [29], peut mener à une amélioration.

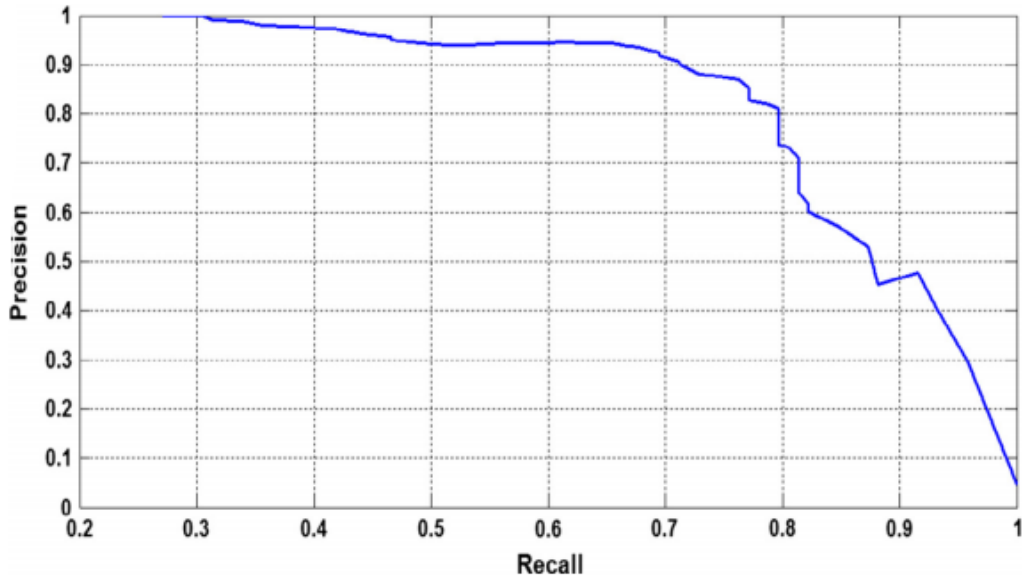

FIGURE 5.6 – Performances de détection de l'objet *"bateau"* en terme de précision et de rappel.

Nous remarquons qu'il existe quelques catégories d'événements qui sont très difficiles à classer, telles que "Birthday", où la plupart de photos des album ne contiennent pas des images avec des objets/scènes pertinents. Notre modèle n'a pas réussi à classer correctement ces albums, qui sont également difficiles à classer par les humains. La figure 5.7 montre l'impact du nombre d'images représentatives sur la précision de la prédiction des classes d'événements. Une image est considérée comme image représentative si elle contient une ou plusieurs scène(s) pertinente(s) et/ou plusieurs objet(s) pertinent(s) caractérisant une classe d'événement particulière.

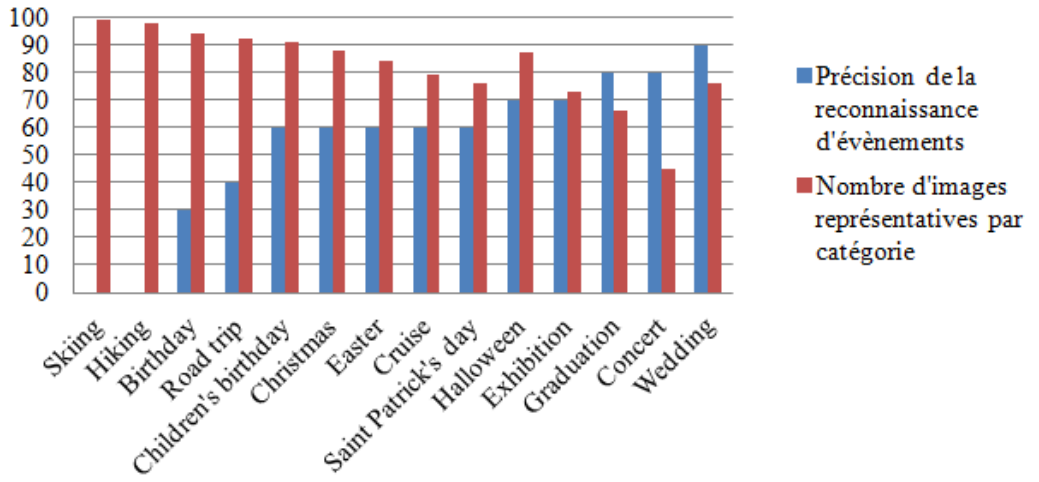

FIGURE 5.7 – Impact du nombre d'images représentatives sur la précision de la prédiction des classes d'événements.

Dans l'ensemble, la précision de la classification augmente progressivement avec le nombre d'images représentatives. Nous remarquons que généralement l'absence de photos représentatives est une conséquence directe du comportement adopté par le photographe lors de la prise des photos. Parfois, l'acquisition de photo est sporadique lorsque l'utilisateur ne se concentre pas sur les objets/scènes principaux. La figure 5.8 montre un exemple d'album "Birthday" résultant de ce comportement sporadique. Les photos de cet album ne contiennent pas assez d'information pour discriminer l'événement. Le dernier problème rencontré est la variabilité et la diversité culturelle de certains contenus des classes. Deux exemples de ce cas sont présentés dans les figures 5.9 et 5.10 qui correspondent à l'événement "Exhibition (exposition)" avec un contenu très diversifié, et un événement de " Japanese wedding (mariage japonais)" avec une mariée et un mari qui ne portent pas de robe de mariée blanche et de costume, respectivement. Quelques images seulement ont une relation directes avec l'événement principal.

5.6 Évaluation de la pertinence des caractéristiques

Pour valider l'avantage d'utiliser la pertinence des caractéristiques pour la reconnaissance d'événements, nous avons mis en œuvre deux autres versions

FIGURE 5.8 – Exemple d'un événement 'd'anniversaire" tiré de la base PEC illustrant le comportement sporadique de prise de photos d'un utilisateur.

de notre modèle. La première version est (OS+PGM) qui combine les objets et les scènes en ignorant la pertinence de ces caractéristiques. La deuxième version, (R+OS+PGM) reflète le modèle complet tel que proposé dans la figure 4.1. Elle combine les objets et les scènes en prenant en considération la pertinence de ces caractéristiques. Les figures 5.11 et 5.12 montrent les

FIGURE 5.9 – Variabilité du contenu d'images dans un événement "exposition".

FIGURE 5.10 – Exemple d'un événement "mariage" mal classé.

matrices de confusion obtenues pour les 14 événements en utilisant les deux versions de notre modèle (OS+PGM) et (R+OS+PGM), respectivement.

Nous remarquons que même sans utiliser la pertinence des caractéristiques objet et scène, la majorité des classes d'événements ont été correctement catégorisées en utilisant le modèle (OS+PGM), où la précision moyenne obtenue est de 70%. Certaines classes d'événements ont atteint une précision très élevée, par exemple "Concert" et "Skiing". Cela résulte probablement de : 1)

l'utilisation d'excellentes caractéristiques qui peuvent discriminer des classes spécifiques, telle que la scène "snowy mountain (montagne enneigée)" pour l'évènement "Skiing" ; 2) le nombre important d'images représentatives dans l'album, le cas dans les albums de "Concert".

La combinaison d'objets et de scènes aide à distinguer entre des événements ayant un contenu visuel similaire et produit des résultats plus fiables par rapport à l'utilisation de scènes et d'objets séparément. Par exemple, dans l'événement "Graduation", les objets améliorent considérablement la précision de la classification. En effet, la caractéristique principale de l'événement "Graduation" est la présence des objets *mortarboard (toge)* et de *gown (robe universitaire)*, quel que soit la scène de l'événement (intérieur, extérieur ou leur combinaison). Dans l'événement "Cruise", les scènes contribuent à améliorer les performances de la catégorisation des événements. Dans ce cas, les gens ont tendance à passer plus de temps à l'extérieur du *boat (bateau)*, et la plupart des photos contiennent des scènes *sea (mer)* et *coast (côte)*.

L'introduction de la pertinence des caractéristiques, par la version (R+OS+PGM), a amélioré les résultats de la combinaison objet et scène de 4.28%. Ceci peut être observé surtout lorsque la précision de certaines classes a légèrement diminué lors de la combinaison des scènes et d'objets. Par exemple, dans les événements "Children's birthday" et "Easter", il existe une différence entre les précisions obtenues à l'aide des caractéristiques individuelles et leur combinaison. L'introduction de la pertinence des caractéristiques a permis d'atténuer le problème en accentuant la contribution des caractéristiques discriminatives et en négligeant la contribution des caractéristiques non informatives, ce qui a permis une meilleure classification des albums. Cela a également un bon impact sur l'amélioration de la précision de classification lorsque les fréquences des scènes et/ou des objets clés sont faibles.

Il n'est pas surprenant que lorsqu'un événement ne peut être correctement classé à l'aide de scènes ou d'objets, que la combinaison des deux ne peut pas donner des bons résultats. De plus, nous avons remarqué que malgré qu'une caractéristique peut être présente dans plusieurs classes d'événements, elle peut aider à réduire l'ensemble de classes d'événements cibles possibles parmi les 14 classes initiales. Par exemple, l'objet "bouteille de vin" peut être présent dans de multiples événements tels que "Christmas", "Birthday"

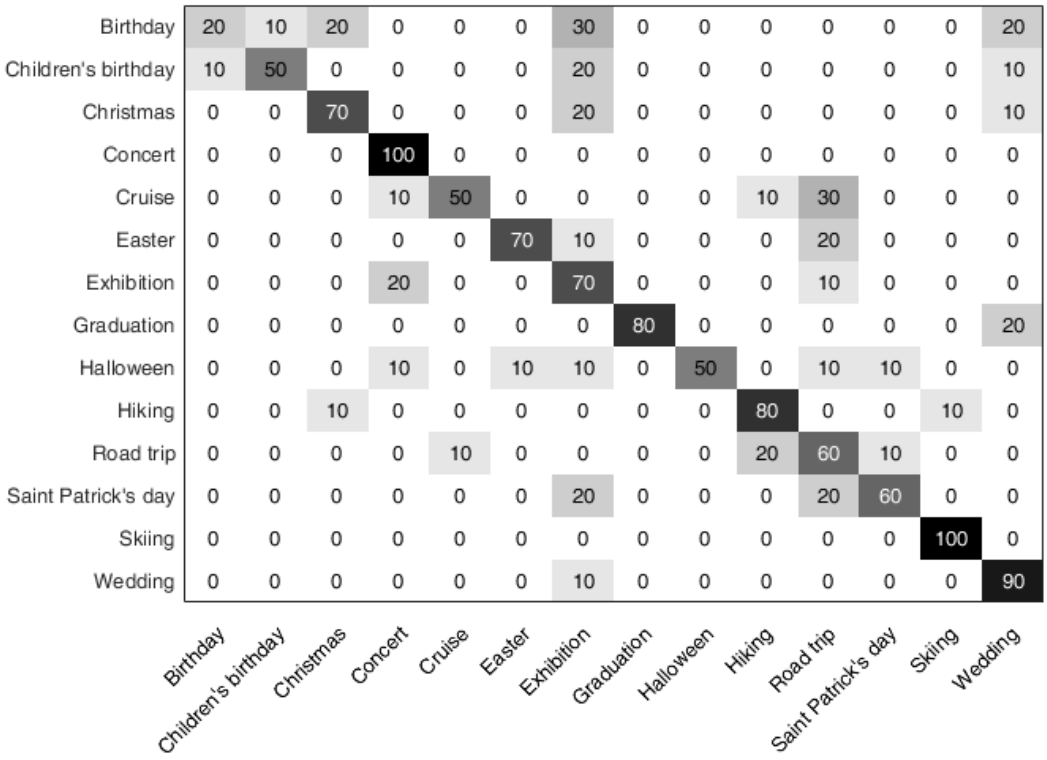

FIGURE 5.11 – Matrice de confusion des 14 événements en combinant les scènes et les objets (OS+PGM), sans utiliser la pertinence de caractéristiques.

et "Wedding", et absent dans "Children's birthday". L'objet "bottle of wine (bouteille de vin)", par conséquent, rétrécit l'ensemble des événements cibles (contenant l'objet) en augmentant leur probabilité maximale.

5.7 Comparaison de notre méthode avec les travaux existants

Dans cette section, nous avons comparé notre approche avec quatre approches de l'état de l'art basées [38, 54, 61, 49]. Dans [49], les auteurs ont présenté des résultats en utilisant différentes configurations de leur méthode. Pour avoir une comparaison équitable, nous avons considéré les meilleurs résultats obtenus par ces différentes configurations. Dans [54], les auteurs ont signalé que la combinaison de plusieurs caractéristiques permet d'obtenir les meilleurs résultats. Par conséquent, pour comparer avec notre travail, nous avons mis en œuvre leur méthode basée sur la combinaison de caractéristiques constituées de : la pyramide spatiale (SPM [35]) en utilisant les descripteurs

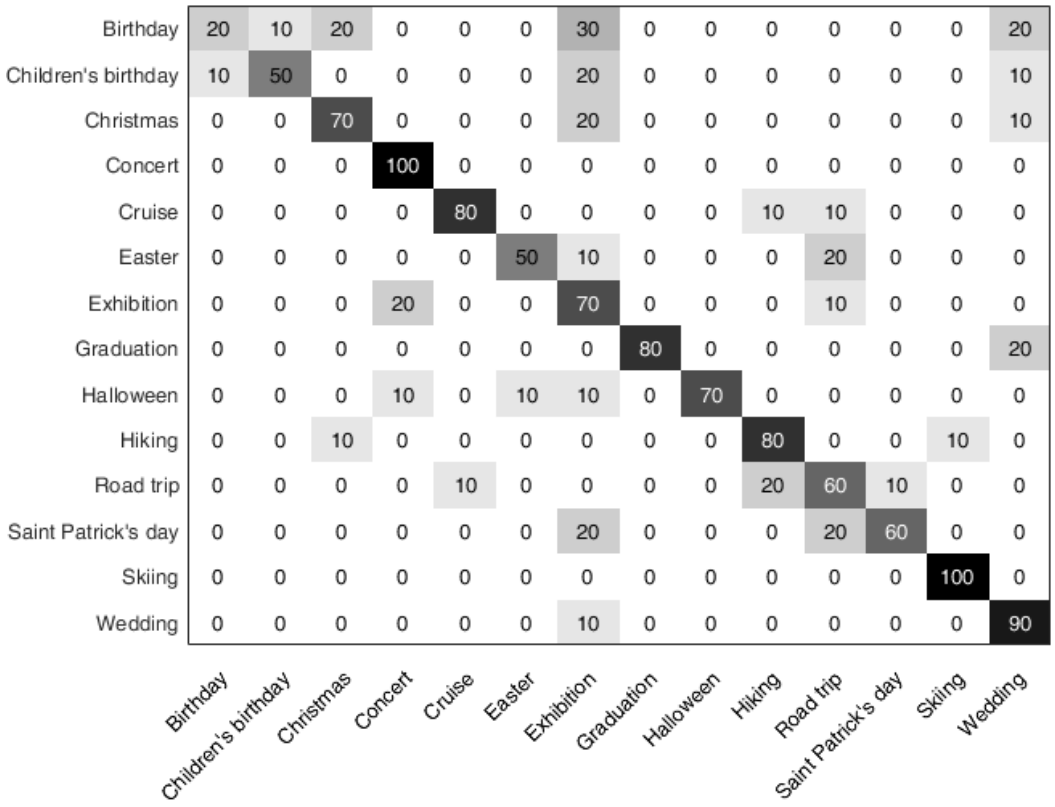

FIGURE 5.12 – Matrice de confusion des 14 événements en combinant les scènes et les objets, en utilisant la pertinence de caractéristiques (R+OS+PGM).

SIFT (Scale Invariant Feature Transform), Couleur et Regularized Max Pooling (RMP [174]). La taille de vocabulaire utilisée pour le sac de mots de SIFT+Couleur est de 2000. Pour l'extraction des caractéristiques CNN, nous avons utilisé l'implémentation de Caffe avec l'architecture CNN disponible publiquement décrit par Krizhevsky et al. [106]. Ceci donne un vecteur de dimension 4096. Le SPM et le RMP sont tous les deux basés sur des machines à vecteurs de support des moindres carrés (Least Squares Support Vector Machines- LSSVM [175]). L'évaluation a été effectuée sur les données PEC et les résultats obtenus sont présentés dans le tableau 5.2.

En termes de précision de classification, notre méthode atteint une moyenne de 74.29%, dépassant les meilleures précisions moyennes obtenues par [49], [54], [61] et [38] correspondant à $0.85\%, 32.57\%, 17.14\%$ et 18.57% de plus, respectivement. Plus précisément, notre méthode permet une classi-

Events \ Methods	Bossard *et al.* [38]	Wu et *al.* [49]	O-PGM	S-PGM	OS-PGM	R-OS-PGM	Tsai *et al.*[61]	Kwon *et al.* [54]
Birthday	10 %	12 %	20 %	30 %	20%	**30** %	0 %	0 %
Children's birthday	30 %	57 %	**60** %	0 %	50%	**60** %	**60** %	10 %
Christmas	70 %	**89**%	60 %	30 %	70%	80%	60 %	40 %
Concert	**100**%	**100**%	80 %	100 %	100%	**100**%	80 %	100 %
Cruise	50%	**82**%	60 %	70 %	80%	80%	70 %	40 %
Easter	50 %	44%	**60** %	0 %	50%	**60**%	**60** %	20%
Exhibition	70%	**75**%	70 %	70 %	70%	70 %	50 %	50%
Graduation	40%	69%	80 %	30 %	80%	**90**%	70 %	40%
Halloween	30%	**82**%	70 %	00 %	70%	70%	70 %	10%
Hiking	**80**%	52%	10 %	**80** %	**80**%	**80**%	40 %	70%
Road trip	40%	**91**%	40 %	50 %	60%	60%	10 %	30%
Saint Patrick's day	30%	**98**%	60 %	10 %	60%	70%	40 %	90%
Skiing	100%	**100**%	0 %	**100** %	**100**%	**100**%	**100** %	60%
Wedding	80%	77%	**90** %	10 %	**90**%	**90**%	**90** %	10%
Average accuracy	55.71%	73.43%	54.28%	41.42%	70%	**74.28**%	57.14 %	41.71%
Average F_1 measure	56.16%	57.68%	55.88 %	41.72%	71.01%	**74.82**%	60.11 %	38.62%

TABLE 5.2 – Comparaison de notre méthode avec avec quelques travaux existants sur la base PEC. Les valeurs en pourcentage représentent la précision moyenne obtenue pour chaque catégorie d'événement. Les meilleurs scores obtenus sont en gras.

fication plus précise que les autres méthodes pour les événements "Birthday, Children's birthday, Easter, Graduation et Wedding". Pour les autres catégories d'événements, nous avons atteint une performance proche des autres méthodes. En termes de F score, nous avons obtenu un score moyen de 74.82%, dépassant les meilleurs F scores obtenus par [49], [54], [61] et [38] avec 17.17%, 36.2%, 14.71% et 18.66% de plus, respectivement. Cela s'explique par le fait que le F score est une mesure de compromis entre la précision et rappel. Autrement dit, la précision obtenue par notre méthode est plus élevée que celle des autres méthodes. Cela confirme, entre autres, que l'intégration de la pertinence des caractéristiques et la combinaison de scènes et d'objets sont plus appropriées pour la classification d'événements.

À partir des résultats obtenus par [54], il est clair que la représentation basée sur des caractéristiques de bas niveau n'est plus suffisante pour obtenir une bonne reconnaissance d'événements dans les collections de photos personnelles. Dans leur travail, l'utilisation de SPM avec les caractéristiques SIFT+couleur produit des descripteurs de $42K$ dimensions, où $K = 2000$ représente le nombre de cases d'histogramme. En ajoutant le RMP aux caractéristiques CNN, les vecteurs de caractéristiques finaux seront d'une dimension de 88096. Cependant, en dépit de la grande taille des vecteurs de caractéristiques, la méthode a un pouvoir discriminatif limité pour reconnaître les catégories d'événements avec un contenu visuel complexe. Notre modèle

(O+PGM) et [61] ont obtenu une performance de classification proche sauf pour l'événement "Skiing". Cela est dû au fait que les scores de détection de tous les objets de l'événement "Skiing", tels que "moto de neige" et "masque de ski", étaient sous le seuil de confiance. Par conséquent, aucun objet n'a été détecté pour cette classe, ce qui a conduit à une précision de 0% en utilisant notre modèle probabiliste d'objet. Cependant, toutes les images avec un vecteur d'objet vide sont classées dans l'événement "Skiing" à l'aide de SVM, ce qui donne une précision de 100% ; toutefois, il donne un mauvais F score (34, 48% pour la catégorie "Skiing").

Enfin, pour réduire le nombre d'images utilisées pour la classification, Wu et al. [49] proposent de tirer aléatoirement de chaque album un nombre d'images. Cependant, la sélection aléatoire ne garantit pas la présence de photos représentatives parmi celles choisies. De plus, l'échantillonnage aléatoire peut conduire à un sur-apprentissage des albums en raison de la redondance du contenu visuel entre les images. L'introduction de la pertinence des caractéristiques permet de renforcer la contribution des caractéristiques discriminatives et, par conséquent, réduire l'effet des images redondantes et non représentatives. En outre, notre méthode donne des résultats plus interprétables sémantiquement car les caractéristiques, constitués d'objets et de scènes, appartiennent à l'espace sémantique et peuvent être utilisées pour indexer les albums. Cela peut être très utile pour plusieurs applications entre autres la recherche des albums photos.

5.8 Temps de calcul

Nous avons évalué le temps de calcul de notre méthode. Puisque les albums sont représentés à l'aide des scènes et des objets, nous ne prenons pas en compte le temps des phases de reconnaissance d'objets et de reconnaissance de scènes, car elles dépendent de l'implémentation de la toolbox Caffe [173]. Nous utilisons les 140 albums, suggérés par le protocole de tests de la base d'images PEC, pour obtenir le temps de traitement. Notre code a été écrit avec Matlab, et il a été exécuté sur un ordinateur doté d'un cœur Intel i7 et 8 GB de RAM. Notre modèle a un temps d'inférence rapide avec une moyenne de 0, 5 secondes pour les 140 albums. Nous remarquons que le temps de calcul est approximativement linéaire par rapport au nombre d'images par album.

Il est clair que le temps de traitement moyen par album dépend du nombre d'images qu'il contient. Le temps de calcul de la méthode proposée pour un album avec un nombre de photos entre 24 et 500 est illustré dans la figure 5.13.

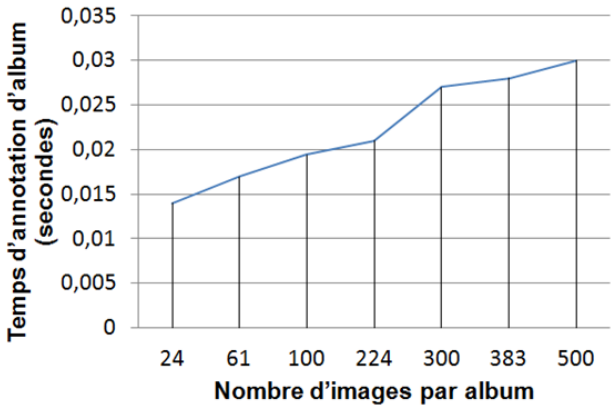

FIGURE 5.13 – Temps de prédiction d'événements en fonction des tailles des albums.

5.9 Conclusion

Nous avons présenté, dans ce chapitre, les expérimentations effectuées pour valider le modèle proposé pour la reconnaissance d'événements dans les albums photos. Durant ces expérimentations, nous avons évalué l'impact de l'utilisation de chacun des objets et des scènes individuellement sur les performances ainsi que l'impact de leur combinaison. Pour évaluer l'utilité de l'introduction de la pertinence des caractéristiques, nous avons implémenté deux versions de notre modèle avec et sans l'introduction de la pertinence. Les résultats obtenus ont montré que l'utilisation de la pertinence des caractéristiques améliore la précision de la reconnaissance d'évènements. Cette notion permet de mettre l'accent sur les informations discriminatives lors de la classification en assurant une reconnaissance d'événements plus précise. Nous avons établi une étude comparative avec d'autres approches existantes. Notre modèle a bien pu surpasser l'état de l'art en donnant les meilleurs résultats.

CONCLUSION

Le développement d'un système d'annotation d'images est devenu de plus en plus important, ceci est dû à la nécessité de faciliter la recherche et la gestion de la quantité phénoménale des images. Les travaux présentés dans cette thèse se situent dans le cadre général de l'annotation d'images et plus particulièrement dans la reconnaissance d'événements dans les albums photos. La classification des albums de photos personnelles représente un vrai challenge aussi bien pour la communauté de vision par ordinateur que pour celle de recherche d'images. Une des difficultés rencontrées dans la classification des albums photos réside dans la diversité du contenu des photos personnelles. Durant ce travail de thèse, nous nous sommes intéressés à l'annotation des albums photos représentant un événement particulier tel que anniversaire, mariage, etc. Afin de mieux représenter la diversité du contenu des photos d'un album prises dans un événement particulier, nous avons proposé une représentation d'album basée sur les caractéristiques de haut niveau. Ces dernières sont constituées de scènes et d'objets extraits de l'ensemble de photos, ce qui n'est pas la modélisation abordée communément dans la littérature. Cette approche permet de projeter l'album photos dans un espace sémantique plus riche.

Dans le premier chapitre, nous avons donné un aperçu général sur le concept de la recherche et l'annotation d'images. Nous avons présenté les systèmes d'annotation d'images avec leur deux types : systèmes de recherche par texte et systèmes de recherche par contenu visuel. Pour chacun de ces types, nous avons présenté leurs avantages et leurs inconvénients. Nous avons mis l'accent sur la technique d'annotation sémantique d'images avec ses différentes méthodes existantes à savoir l'annotation manuelle et automatique. Aussi, nous avons établi un état de l'art des méthodes de reconnaissance d'événements dans les images. Nous avons clôturé le chapitre par une synthèse des travaux existants et nous avons situé notre contribution par rapport à l'existant.

Dans le deuxième chapitre, nous avons introduit les différents types d'apprentissage d'une base d'images en l'occurrence les méthodes d'apprentissage supervisé et les d'apprentissage non supervisé. Nous avons présenté quelques méthodes discriminatives telles que kppv, les SVMs et les réseaux de neurones

pour l'apprentissage supervisé. Nous avons également parcouru les différentes approches de sélection de caractéristiques pertinentes pour la classification.

Le troisième chapitre est dédié à la présentation détaillée des réseaux bayésiens ainsi que leurs propriétés. Il s'agit, en fait, de l'apprentissage de la structure et les paramètres des réseaux bayésiens. Nous avons aussi détaillé le principe d'inférence bayésienne.

Dans le quatrième chapitre, nous avons présenté notre réseau bayésien pour la reconnaissance d'événements dans un album photos. En commençant par l'extraction des caractéristiques sémantiques en utilisant les réseaux de neurones à convolution. Par la suite, nous avons détaillé la structure de notre modèle, l'apprentissage de ses paramètres à savoir : les paramètres d'objet, de scène et d'événement ainsi que les paramètre de pertinence de caractéristiques, et enfin, la prédiction de classe d'événement d'un nouvel album par l'inférence bayésienne .

L'évaluation de notre modèle, présentée au cinquième chapitre, indique que la combinaison de caractéristiques sémantiques, objets et scènes, et l'incorporation de leur pouvoir discriminatif dans l'inférence, aide à combler le fossé sémantique dans les albums photos. L'efficacité de notre modèle a été mise en évidence en le comparant avec d'autres méthodes d'état de l'art.

Pour conclure, les albums photos peuvent être caractérisés par plusieurs types d'attributs : visuels, sémantiques et contextuels. Notre hypothèse est que pour obtenir de bons résultats pour la prédiction d'événements dans les albums photos, il faut représenter les albums par des caractéristiques de haut-niveau et inclure leur pouvoir discriminatif dans la classification. Par conséquent, nous avons opté pour le développement d'un modèle probabiliste qui combine les différents caractéristiques sémantiques, objets et scènes, de toutes les images d'un albums. Par ailleurs, notre modèle intègre la notion de pertinence de caractéristiques.

Le domaine de la reconnaissance d'événements dans les albums photos étant très vaste, il reste encore beaucoup d'efforts à fournir afin de développer des outils efficaces. Parmi les pistes prometteuses qu'il faudrait explorer, il y a l'incorporation des informations contextuelles, telles que les informations spatiotemporelles, pour la prédiction des événements. En effet, si on arrive à bien exploiter les informations contextuelles, cela devrait contribuer de ma-

nière efficace à l'amélioration de la précision de reconnaissance d'événement au contenu variables. Une autre piste qui semble prometteuse est l'utilisation de modèles de reconnaissance d'objets et de reconnaissance de scènes plus performants. Nous avons remarqué durant les expérimentations que les scènes et les objets non détectés augmentent le taux de classification erroné. L'exploitation de meilleurs modèles de CNN, ou de nouvelle méthodes pour une reconnaissance de scènes et d'objets plus précise peut aider à pallier à ce problème. Cela permettra d'améliorer la qualité de prédiction des événements en ayant une représentation d'albums plus correcte.

REFERENCES

1. S. Ren K. He, X. Zhang and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

2. B. Mesman M. Peemen and H. Corporaal. Efficiency optimization of trainable feature extractors for a consumer platform. 6915 :293–304, 08 2011.

3. L. J. Li and L. Fei-Fei. What, where and who ? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

4. B. Rand. *The Classical Psychologists : Selections Illustrating Psychology.* Anaxagoras to Wundt, Houghton, Mifflin and Company, Boston, MA, US, 1912.

5. J. L. Marignier. *Aux origines de la photographie : Nicéphore Niépce. Séance de l'Académie des beaux-arts.* goo.gl/puCGeX, 2008.

6. M. Mullaney. *Digital camera pioneer StevenSasson to receiveRensselaer Polytechnic Institute 2011 daviesmedal.* School of Engineering Rensselaer, http ://eng.rpi.edu/news/09202011-2000/digital-camera-pioneer-steven-sasson-receive-rensselaer-polytechnic-institute.

7. Site web : http://www.info-histoire.com/15658/premieres-photos-histoire/. Visité le : 31/10/2017.

8. J. Tesic. Metadata practices for consumer photos. *IEEE MultiMedia*, 12(3) :86–92, July 2005.

9. A. C. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Transactions on Multimedia*, 5(3) :390–402, Sept 2003.

10. G. Quellec. *Indexation et fusion multimodale pour la recherche d'information par le contenu. Appliquation aux bases de donnée d'images médicale.* PhD thesis, Université européenne de Bretagne, France, 2008.

11. A. B. Tucker. *Computer Science Handbook, Second Edition.* Chapman & Hall/CRC, 2004.

12. G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, HUC '99, pages 304–307, London, UK, UK, 1999. Springer-Verlag.

13. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1349–1380, Dec 2000.

14. S. Ayache. *Indexation de documents vidéos par concepts et par fusion de caractéristiques audio, image et textel.* PhD thesis, Institut Polytechnique de Grenoble (INPG), France, 2007.

15. N. Hervé. *Vers une description efficace du contenu visuel pour l'annotation automatique d'images.* PhD thesis, Université d'Orsay, Paris-Sud 11, France, 2009.

16. Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, 40(1) :262–282, 2007.

17. J. S. Hare, P. H. Lewis, P. GB. Enser, and C. J. Sandom. Mind the gap : Another look at the problem of the semantic gap in image retrieval. *Proc.SPIE*, 2006.

18. S. Bacha and N. Benblidia. Combining context and content for automatic image annotation on mobile phones. In *2013 International Conference on IT Convergence and Security (ICITCS)*, pages 1–4, Dec 2013.

19. C. Qin, X. Bao, R. Roy Choudhury, and S. Nelakuditi. Tagsense : A smartphone-based approach to automatic image tagging. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, pages 1–14, New York, NY, USA, 2011. ACM.

20. Y. S Lee and S. B Cho. A mobile picture tagging system using treestructured layered bayesian networks. *Mobile Information Systems*, 9(3) :209–224, 2013.

21. O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5) :1055–1064, Sep 1999.

22. A. Makadia, V. Pavlovic, and S. Kumar. *A New Baseline for Image Annotation*, pages 316–329. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

23. H. Bouyerbou, S. Oukid, N. Benblidia, and K. Bechkoum. Hybrid image representation methods for automatic image annotation : A survey. In *2012 International Conference on Signals and Electronic Systems (ICSES)*, pages 1–6, Sept 2012.

24. D. Kim and H. Park. *An Efficient Face Recognition through Combining Local Features and Statistical Feature Extraction*, pages 456–466. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

25. Y. Su, S. Shan, X. Chen, and W. Gao. Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Transactions on Image Processing*, 18(8) :1885–1896, Aug 2009.

26. R. S. Kathavarayan and M. Karuppasamy. Preserving global and local features for robust face recognition under various noisy environments. *International Journal of Image Processing (IJIP)*, 3(6) :328, 2010.

27. N. Devarajan S. Anila. Global and local classifiers for face recognition. *European Journal of Scientific Research*, 57(4) :556–566, 2011.

28. D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, and M. C. Benfield. Combining local and global image features for object class recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 47–47, June 2005.

29. L. Wang, Z. Wang, W. Du, and Y. Qiao. Object-scene convolutional neural networks for event recognition in images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 30–35, June 2015.

30. I. H. Witten, E. Frank, and M. A. Hall. *Data Mining : Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

31. L. Cao, J. Luo, H. Kautz, and T. S. Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *IEEE Transactions on Multimedia*, 11(2) :208–219, Feb 2009.

32. R. Szeliski. *Computer Vision, Algorithms and Applications*, page 612. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

33. D. Larlus. *Création et utilisation de vocabulaires visuels pour la catégorisation d'images et la segmentation de classes d'objets.* PhD thesis, Institut National Polytechnique de Grenoble (INPG), France, 2008.

34. Li. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 524–531, Washington, DC, USA, 2005.

35. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision & Pattern Recognition (CPRV '06)*, pages 2169 – 2178, New York, United States, June 2006. IEEE Computer Society.

36. A. V. Manjaly and B. S. Priya. Malayalam text and non-text classification of natural scene images based on multiple instance learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pages 190–196, 2016.

37. J. Vogel, A. Schwaninger, C. Wallraven, and H. H. Bülthoff. Categorization of natural scenes : Local versus global information and the role of color. *ACM Trans. Appl. Percept.*, 4(3), November 2007.

38. L. Bossard, M. Guillaumin, and L. Van. Event recognition in photo collections with a stopwatch hmm. In *2013 IEEE International Conference on Computer Vision*, pages 1193–1200, Dec 2013.

39. M. Brenner and E. Izquierdo. *Joint People Recognition across Photo Collections Using Sparse Markov Random Fields*, pages 340–352. Springer International Publishing, Cham, 2014.

40. W. W. Y. Ng, T. M. Zheng, P. P. K. Chan, and D. S. Yeung. Social relationship discovery and face annotation in personal photo collection. In *2011 International Conference on Machine Learning and Cybernetics*, volume 2, pages 631–637, July 2011.

41. N. O'Hare and A. F. Smeaton. Context-aware person identification in personal photo collections. *IEEE Transactions on Multimedia*, 11(2) :220–228, Feb 2009.

42. A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the 2Nd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '02, pages 326–335, New York, NY, USA, 2002. ACM.

43. S. Kisilevich, D. Keim, N. Andrienko, and G. Andrienko. *Towards Acquisition of Semantics of Places and Events by Multi-perspective Analysis of Geotagged Photo Collections*, pages 211–233. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

44. M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.*, pages 53–62, June 2004.

45. A. Ulges, M. Worring, and T. Breuel. Learning visual contexts for image annotation from flickr groups. *IEEE Transactions on Multimedia*, 13(2) :330–341, April 2011.

46. C. Zigkolis, S. Papadopoulos, G. Filippou, Y. Kompatsiaris, and A. Vakali. Collaborative event annotation in tagged photo collections. *Multimedia Tools and Applications*, 70(1) :89–118, 2014.

47. W. Jiang and A. C. Loui. Semantic event detection for consumer photo and video collections. In *2008 IEEE International Conference on Multimedia and Expo*, pages 313–316, June 2008.

48. N. Imran, J. Liu, J. Luo, and M. Shah. Event recognition from photo collections via pagerank. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 621–624, New York, NY, USA, 2009. ACM.

49. Z. Wu, Y. Huang, and L. Wang. Learning representative deep features for image set analysis. *IEEE Transactions on Multimedia*, 17(11) :1960–1968, Nov 2015.

50. R. Rothe, R. Timofte, and L. V. Gool. Dldr : Deep linear discriminative retrieval for cultural event classification from a single image. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 295–302, Dec 2015.

51. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

52. R. Timofte and L. Van Gool. Iterative nearest neighbors for classification and dimensionality reduction. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2456–2463, June 2012.

53. M. S. Dao, D. T. Dang Nguyen, and F. G. B. De Natale. Robust event discovery from photo collections using signature image bases (sibs). *Multimedia Tools and Applications*, 70(1) :25–53, 2014.

54. H. Kwon, K. Yun, M. Hoai, and D. Samaras. Recognizing cultural events in images : A study of image categorization models. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 51–57, June 2015.

55. F. Tang, D. R. Tretter, and C. Willis. Event classification for personal photo collections. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 877–880, May 2011.

56. A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, and X. Girói Nieto. Cultural event recognition with visual convnets and temporal models. *CoRR*, abs/1504.06567, 2015.

57. J. Yuan, J. Luo, and Y. Wu. Mining compositional features from gps and visual cues for event recognition in photo collections. *IEEE Transactions on Multimedia*, 12(7) :705–716, Nov 2010.

58. S. Lefèvre. *Détection d'évènements dans une séquence vidéo*. PhD thesis, Université François Rabelais, France, 2002.

59. R. El Ouazzani. *La reconnaissance et l'apprentissage des évènements chauds dans la vidéo de matches de football en utilisant les Modèles de Markov Cachés*. PhD thesis, Université Mohammed V, Maroc, 2010.

60. C. Sun and R. Nevatia. Active : Activity concept transitions in video event classification. In *2013 IEEE International Conference on Computer Vision*, pages 913–920, Dec 2013.

61. S.-F. Tsai, T. S. Huang, and F. Tang. Album-based object-centric event recognition. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6, July 2011.

62. S.-F. Tsai, L. Cao, F. Tang, and T. S. Huang. Compositional object pattern : A new model for album event recognition. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 1361–1364, New York, NY, USA, 2011. ACM.

63. M. Liu, X. Liu, Y. Li, X. Chen, A. G. Hauptmann, and S. Shan. Exploiting feature hierarchies with convolutional neural networks for cultural event recognition. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 274–279, Dec 2015.

64. Handbook of categorization in cognitive science.

65. C. Galleguillos and S. Belongie. Context based object categorization : A critical survey. *Comput. Vis. Image Underst.*, 114(6) :712–722, June 2010.

66. K. M. Swallow, J. M. Zacks, and R. A. Abrams. Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology*, 138(2) :236, 2009.

67. J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. Event perception : A mind-brain perspective. *Psychological Bulletin*, 133(2) :273–293, 2007.

68. M. Dyck and M. B. Brodeur. {ERP} evidence for the influence of scene context on the recognition of ambiguous and unambiguous objects. *Neuropsychologia*, 72 :43 – 51, 2015.

69. L. Mudrik, D. Lamy, and L. Y. Deouell. {ERP} evidence for context congruity effects during simultaneous object-scene processing. *Neuropsychologia*, 48(2) :507 – 517, 2010.

70. J.S. Bowers and C.J. Davis. Bayesian just-so stories in psychology and neuroscience. 138(3) :389–414, 2012.

71. W. S. Geisler and R. L. Diehl. A bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, 27 :379–402, 2003.

72. W. J. Ma. Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, 16(10) :511 – 518, 2012.

73. P. Vincent. *Modèles à noyaux à structure locale.* PhD thesis, Université de Montréal, Canada, 2003.

74. L. Banabbou. *Contributions à la classification supervisée multi-classes et multicritères en aide à la décision.* PhD thesis, Université de Laval, Canada, 2009.

75. S. J. Delany and P. Cunningham. An analysis of case-base editing in a spam filtering system. *Advances in Case-Based Reasoning*, pages 3–25, 2004.

76. G. Guo, S. Z. Li, and K. Chan. Face recognition by support vector machines. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 196–201. IEEE, 2000.

77. Q. Gao, P. Forster, K. R. Mobus, and G. S. Moschytz. Fingerprint recognition using cnns : Fingerprint preprocessing. In *Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on*, volume 3, pages 433–436. IEEE, 2001.

78. O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 4277–4280. IEEE, 2012.

79. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016.

80. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

81. Ji-Hyun Kim. Estimating classification error rate : Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11) :3735–3745, 2009.

82. Andrew Y. Ng. Preventing "overfitting" of cross-validation data. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 245–253. Morgan Kaufmann, 1997.

83. J. Hughes-Oliver, W. J. Welch, and H. Shen. Efficient, adaptive cross-validation for tuning and comparing models, with application to drug discovery. *The Annals of Applied Statistics*, 5(4) :2668–2687, 2011.

84. V. Vandewalle. *Estimation et sélection en classification semi-supervisée*. PhD thesis, Université Lille1, France, 2009.

85. Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. Unsupervised supervised learning i : Estimating classification and regression errors without labels. *J. Mach. Learn. Res.*, 11 :1323–1351, August 2010.

86. A. Hamadi. *Utilisation du contexte pour l'indexation sémantique des images et vidéos*. PhD thesis, Université de Grenoble, France, 2006.

87. G. Bouchard. *Les modèles génértifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle*. PhD thesis, Université Joseph Fourier - Grenoble 1, France, 2005.

88. R. Rosales and S. Sclaroff. Combining generative and discriminative models in a framework for articulated pose estimation. *International Journal of Computer Vision*, 67(3) :251–276, May 2006.

89. Z. Tu. Learning generative models via discriminative approaches. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

90. K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

91. B.-V. Dasarathy. Nearest neighbor ({NN}) norms :{NN} pattern classification techniques. 1991.

92. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. *When Is "Nearest Neighbor" Meaningful ?*, pages 217–235. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

93. C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3) :273–297, September 1995.

94. G. Madzarov, D. Gjorgjevikj, and I. Chorbev. A multi-class svm classifier utilizing binary decision tree. *Informatica*, 33 :225–233, 01 2009.

95. Y. J. Lee and S. Y. Huang. Reduced support vector machines : A statistical theory. *IEEE Transactions on Neural Networks*, 18(1) :1–13, Jan 2007.

96. A. Djeffal. *Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données*. PhD thesis, Université de Biskra, France, 2012.

97. H.-T. Lin and C.-J. Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Technical report, 2003.

98. L. H. Hamel. *Knowledge Discovery with Support Vector Machines*. Wiley-Interscience, New York, NY, USA, 2009.

99. H. Joutsijoki and M. Juhola. *Comparing the One-vs-One and One-vs-All Methods in Benthic Macroinvertebrate Image Classification*, pages 399–413. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

100. M. Baccouche. *Apprentissage Neuronal De caractétistiques Spatio-tempotelles pour la Classifiction Automatique de Séquence Vidéo*. PhD thesis, Université de Lyon, France, 2013.

101. A. K. Jain, Jianchang Mao, and K. M. Mohiuddin. Artificial neural networks : a tutorial. *Computer*, 29(3) :31–44, Mar 1996.

102. F. Rosemblat. *The perceptron : A perceiving and recognizing automation, Rapport technique*. Cornell Aeronautical Laboratory, Ithaca, New York, 1957.

103. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

104. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

105. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, Nov 1998.

106. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012.

107. M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. pages 818–833, 2014.

108. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

109. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

110. S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon. Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1057–1061, Sept 2016.

111. W. Ge and Y. Yu. Borrowing treasures from the wealthy : Deep transfer learning through selective joint fine-tuning. *arXiv preprint arXiv :1702.08690*, 2017.

112. A. K. Reyes, J. C. Caicedo, and J. E. Camargo. Fine-tuning deep convolutional networks for plant recognition. In *CLEF (Working Notes)*, 2015.

113. J. Kazama and J. Tsujii. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 137–144, Stroudsburg, PA, USA, 2003.

114. H. Chouaib. *Sélection de caractéristiques méthodes et applications*. PhD thesis, Université Paris Descartes, France, 2011.

115. G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine learning : proceedings of the eleventh international*, pages 121–129. Morgan Kaufmann, 1994.

116. G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1) :16–28, 2014.

117. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3 :1157–1182, March 2003.

118. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1) :273 – 324, 1997.

119. T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. *Embedded Methods*, pages 137–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

120. T. Hastie, S. Rosset, J. Zhu, and H. Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3) :349–360, 2009.

121. A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press, 2002.

122. J. Louradour. *Noyeaux de séquences pour la vérification du locuteur par Machines à Vecteurs de Support*. PhD thesis, Université de Toulouse III, France, 2007.

123. F. V. Jensen. *Introduction to Bayesian Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 1996.

124. M. D. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine Learning*, 29(2) :181–212, Nov 1997.

125. J. Gross and J. Yellen. *Handbook of Graph Theory*. CRC Press, 2003.

126. Swift R. Rao, M. *Probability Theory with Applications*. Springer US, 2006.

127. J. Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

128. Lauritzen, s. l. and spiegelhalter, d. j. chapter Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems, pages 157–224. JSTOR, 1988.

129. P Naim, P. H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux bayésiens, 3 edition*. Eyrolles, Paris, 2007.

130. S. Barrat and S. Tabbone. Classification and automatic annotation extension of images using a bayesian network. *Traitement du signal*, 26, 2009.

131. O. François. *De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplétes*. PhD thesis, Institut National des Sciences Appliquées de Rouen, 2007.

132. A. Djebbar-Zaidi. *Optimisation de la recherche d'un cas Bayésien*. PhD thesis, Institut National des Sciences Appliquées de Rouen, 2013.

133. S. Li and B. Wang. A method for hybrid bayesian network structure learning from massive data using mapreduce. In *2017 ieee 3rd international conference on big data security on cloud (bigdatasecurity), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids)*, pages 272–276, May 2017.

134. J. Pearl and T. S. Verma. A theory of inferred causation. *Principles of Knowledge Representation and Reasoning*, pages 441–452, 1991.

135. P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Springer-Verlag New York, 1993.

136. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, 2nd Edition*. 2000.

137. D. Maxwell Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of incomplete data given a bayesian network. *CoRR*, abs/1302.3567, 2013.

138. D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks : The combination of knowledge and statistical data. *Machine Learning*, 20(3) :197–243, Sep 1995.

139. G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4) :309–347, Oct 1992.

140. J. M. Pena, R. Nilsson, J. Bjorkegren, and J. Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2) :211 – 232, 2007.

141. S. Rodrigues de Morais and A. Aussem. *A Novel Scalable and Data Efficient Feature Subset Selection Algorithm*, pages 298–312. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

142. I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1) :31–78, Oct 2006.

143. D. M. Chickering. *Learning Bayesian Networks is NP-Complete*, pages 121–130. Springer New York, New York, NY, 1996.

144. N. Friedman, I. Nachman, and D. Peér. Learning bayesian network structure from massive datasets : the sparse candidate algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc., 1999.

145. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2) :131–163, Nov 1997.

146. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression : a statistical view of boosting. *Annals of Statistics*, 28 :2000, 1998.

147. A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41 :1–31, 1979.

148. T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53 :370–418, 1763.

149. F. W. Scholz. *Maximum Likelihood Estimation*. Encyclopedia of Statistical Sciences, John Wiley Sons, Inc., 2004.

150. X. L. Meng and D. Rubin. Maximum likelihood estimation via the ecm algorithm : A general framework. *Biometrika*, 80, 06 1993.

151. A. Gelman. Prior distribution. *Encyclopedia of environmetrics*, 2002.

152. H. Raiffa and R. Schaifer. *Applied statical decision theory, Technical report*. Havard University, 1961.

153. D. Ormoneit and V. Tresp. Improved gaussian mixture density estimates using bayesian penalty terms and network averaging. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, pages 542–548, Cambridge, MA, USA, 1995. MIT Press.

154. E. Parent and J. Bernier. *Le raisonnement bayésien : modélisation et inférence*. Springer Science & Business Media, 2007.

155. N. Shulman and M. Feder. The uniform distribution as a universal prior. *IEEE Transactions on Information Theory*, 50(6) :1356–1362, June 2004.

156. Q. Y. Zhu. *Modèles bayésiens et application à l'estimation des caractéristiques de produits finis et au contrle de la qualité*. PhD thesis, Ecole Nationale des Ponts et Chaussées, France, 1991.

157. B. S. Clarke and A. R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1) :37 – 60, 1994.

158. S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96(453) :270–281, 2001.

159. J. Hasenauer. *Modeling and parameter estimation for heterogeneous cell populations*. Logos Verlag Berlin, 2013.

160. J. L. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298, Apr 1994.

161. C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

162. K. P. Murphy. Binomial and multinomial distributions. *University of British Columbia, Tech. Rep.*, 2006.

163. J. Deng, W. Dong, R. Socher, L. j. Li, K. Li, and L. Fei-Fei. Imagenet : A large-scale hierarchical image database. In *In CVPR*, 2009.

164. D. Bellot. *Fusion de donnÃ©es avec des réseaux bayésiens pour la modèlisation des systèmes dynamiques et son application en télémédecine.* PhD thesis, Université Henri Poincaré - Nancy I, 2002.

165. D. Kasper, G. Weidl, T. Dang, G. Breuel, A. Tamke, A. Wedel, and W. Rosenstiel. Object-oriented bayesian networks for detection of lane change maneuvers. *IEEE Intelligent Transportation Systems Magazine*, 4(3) :19–31, 2012.

166. L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6) :1191–1253, June 2003.

167. S. Bacha, M. S. Allili, and N. Benblidia. Event recognition in photo albums using probabilistic graphical models and feature relevance. *Journal of Visual Communication and Image Representation*, 40(Part B) :546 – 558, 2016.

168. C.P. Robert. *The Bayesian Choice.* Springer-Verlag New York, 2007.

169. J. Azé. *Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs.* PhD thesis, Université de Paris Sud, France, 2012.

170. J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

171. Y. Sasaki. The truth of the f-measure. *Teach Tutor mater*, 1(5), 2007.

172. B. Moulahi. *Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information.* PhD thesis, Université de Toulouse, France, 2016.

173. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe : Convolutional architecture for

fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

174. M. Hoai. Regularized max pooling for image categorization. In *British Machine Vision Conference*, 2014.

175. J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Process. Lett.*, 9(3) :293–300, June 1999.