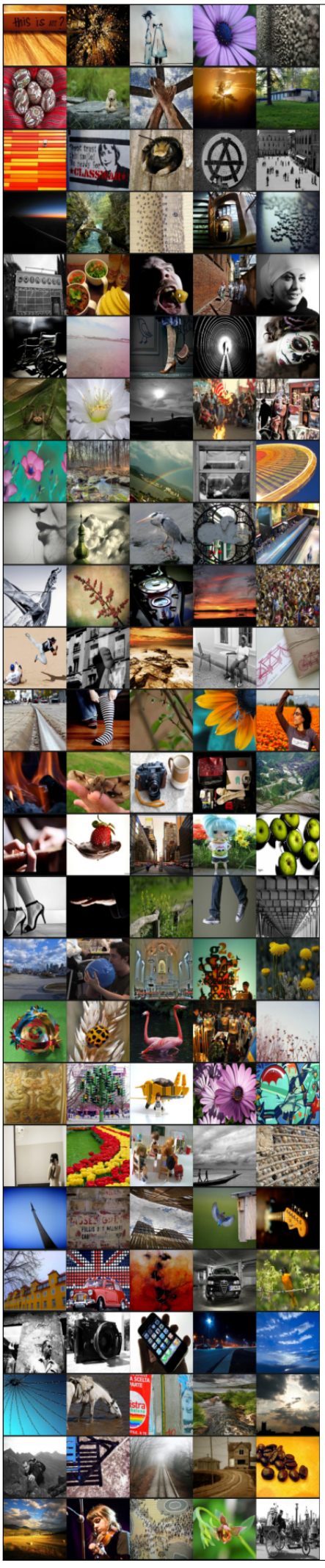


Rapport de Thèse

Apprentissage de Modèles pour la Classification et la Recherche d'Images

Learning Image Classification and Retrieval Models

Thomas Mensink

The research described in this thesis was carried out at the LEAR team of INRIA-Grenoble and the TVPA team of Xerox Research Centre Europe

This work was supported by a CIFRE convention of the ANRT and the French Ministry of Higher Education and Research.

© 2012, T.E.J. Mensink, all rights reserved.

Report de Thèse

1.1 Introduction

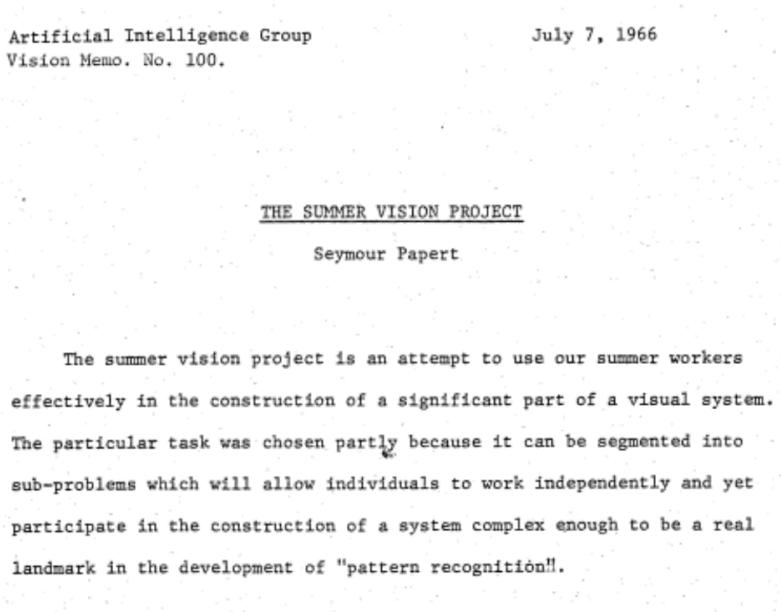
Photos et vidéos sont constamment prises, partagées, regardées, et recherchées. Ceci est la conséquence de la popularisation d'appareils de haute qualité permettant d'acquérir des images (appareils photos numériques et téléphones portables) combiné à la démocratisation des réseaux sociaux et des sites de partage de photos. L'omniprésence des photos et des vidéos sur internet devient évidente lorsqu'on considère les faits suivants:

- Le site de partage de photos Flickr héberge plus de 6 milliards de photos
- Les utilisateurs de Facebook y transfèrent 300 millions de nouvelles photos de façon quotidienne
- Le site de vidéo Youtube présente plus de 3 milliards d'heures de vidéo chaque mois

Ces données visuelles sont souvent accompagnées d'une description ou de méta-données. Ces descriptions se présentent sous différentes formes. Elles peuvent être générées automatiquement, comme les données EXIF, ou la localisation GPS. Elles peuvent être plus riches et haut-niveau, car fournies par les utilisateurs, comme par exemple des légendes, des étiquettes (ou tag). Ces données peuvent prendre la forme de l'insertion d'une image dans une hiérarchie de connaissance comme Wikipedia.

Les méthodes d'interprétation d'images peuvent ajouter de la valeur à ces données visuelles, en permettant notamment de trouver, chercher et utiliser ces images d'une façon sémantiquement pertinente, et centrée sur l'utilisateur, en utilisant le contenu visuel.

Pour illustrer ces possibilités, nous listons quelques applications qui utilisent l'interprétation d'images. Tout d'abord en robotique, la reconnaissance automatique d'objets peut être utilisée lorsqu'un robot doit effectuer la tâche d'attraper un objet spécifique sur une table, tout en se déplaçant autour des meubles dans une pièce. Deuxièmement, dans la recherche d'image, la reconnaissance automatique d'actions permettrait de retrouver certaines scènes de film spécifiques, en se basant sur les actions effectuées par les acteurs. Troisièmement, lors de la création d'un livre photo, un utilisateur pourrait être assisté lors



Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Figure 1.1 *Description d'un projet d'été en vision, «**The summer vision project**» par Seymour Papert, au MIT en 1966. Le but de ce projet était de développer, pendant l'été, un système de reconnaissance visuelle.*

de la composition d'une livre à partir d'une collection de photos, en tenant compte de leur diversité, de leur mémorabilité, et de leur qualité esthétique.

Ces exemples d'applications, ainsi que l'explosion des données visuelles illustrent notre intérêt pour la compréhension d'images, soit dans le cadre de services web à grande échelle, soit à une plus petite échelle, pour aider les utilisateurs à organiser leur collections personnelles de photos.

Dans cette thèse, nous nous focalisons sur l'interprétation des images, et plus précisément sur l'apprentissage supervisé de modèles statistiques pour la classification et la recherche d'images.

1.2 Contexte

À l'aube de l'intelligence artificielle, résoudre le problème de la vision par ordinateur semblait une tâche relativement simple. Ceci est illustré par le projet d'été «*The summer vision project*», proposé par Seymour Papert, au MIT en 1966, voir Figure 1.1. Le but de ce projet était la création d'une bonne partie d'un système visuel, par un large groupe d'étudiants. Ce système de vision artificielle aurait pu être utilisé comme entrée pour des tâches cognitives de haut niveau, comme le raisonnement et la planification. Il était donc un composant nécessaire pour reproduire l'intelligence humaine, ou pour créer un robot intelligent.

Loved the colour of the sky and water against the dark tree reflections.

A long way to go to the top... Just above my house, close to Grenoble, in front of Belledonne mountains.

Figure 1.2 *Exemples de légendes d'images produites par des humains. Les images sont extraites de la base de photos de "Stony Brooks University Captioned Photo Dataset" (Ordonez et al., 2011)*

Quelques dizaines d'années plus tard, il s'avère qu'une compréhension complète du système visuel humain est loin d'être atteinte. La recherche en vision par ordinateur n'est toujours pas capable de reproduire le niveau de compréhension des images qu'a un enfant de 2 ans (Szeliski, 2011).

Cette difficulté est partiellement expliquée par le fait que l'interprétation de scènes et la reconnaissance d'objets sont des problèmes inverses, à travers lesquels nous cherchons à construire une compréhension du monde, étant donné une ou plusieurs images. Il est surprenant de voir que les humains réalisent cette tâche avec si peu d'effort, et avec une telle précision, malgré sa complexité. Le monde visuel dans toute sa complexité est difficile à modéliser, à cause de la quantité énorme de concepts différents, le nombre infini de projection possibles entre une scène en 3 dimensions et une image plan en 2 dimensions, et la complexité des scènes qui peuvent contenir plusieurs niveaux d'occultation et des conditions d'illumination différentes. De plus, il existe une variabilité intrinsèque très grande entre différents objets de la même classe.

1.3 Objectifs

Dans cette thèse, nous nous intéressons au problème de l'interprétation du contenu visuel d'une façon pertinente sémantiquement. Le but ultime serait de créer un système qui puisse générer une description de l'image de la qualité de celle d'un être humain, décrivant les scènes, les objets, et les relations entre ces objets, (voir les deux exemples proposés Figure 1.2).

Un tel système devrait être capable de lier les représentations bas-niveau des images, aux représentations haut-niveau des images, par une description de son contenu. Le manque d'alignement entre les descriptions bas-niveau et l'interprétation haut-niveau des images est souvent connu comme le problème du faussé sémantique (Smeulders et al., 2000).

Dans cette thèse, nous nous intéressons plus particulièrement à l'apprentissage de modèles statistiques pour la classification et la recherche d'images. Ces deux tâches ont pour but de relier des descriptions bas-niveau d'images à une similarité sémantique avec une étiquette de classe, ou avec une autre image, tirant parti des bases d'images étiquetées.

Nous définissons brièvement ces deux tâches ci-dessous:

Classification d'images Le but de la classification d'images est de prédire la pertinence d'un ou plusieurs concepts sémantiques à partir d'un vocabulaire prédéfini. Nous distinguons la classification multi-classe, où les images sont associées à une seule étiquette sémantique, et la classification multi-étiquettes, où une image peut être associée à plus d'une étiquette. La diversité des concepts utilisés dans la classification d'images sont illustrés dans la Figure 1.3.

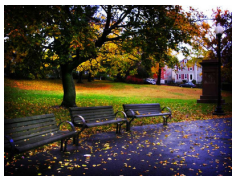
Recherche d'images Le but de la recherche d'images est de retrouver, étant donnée une requête, les images les plus pertinentes d'une base. Il y a deux paradigmes classiques pour la recherche d'images: dans la *requête par texte*, la requête consiste en un ou plusieurs mots-clés, alors que dans la *requête par l'exemple*, la requête est une image, voir la Figure 1.4 pour une illustration.

Ces deux tâches de compréhension d'images sont intiment liées. D'une part, les termes pertinents associés à une image par un système de classification d'image pourraient être utilisés dans un système de recherche basé sur une requête par le texte. D'autre part, la classification d'image peut être vue comme une tâche de recherche par l'exemple, pour laquelle l'image requête est annotée par les étiquettes de ses voisins visuels les plus proches.

Dans cette thèse, nous considérons cinq buts afin d'améliorer les techniques existantes pour ces tâches d'interprétation d'images.

1. **Utilisation de la multi-modalité des données** De grandes bases de données visuelles et multimodales sont actuellement disponibles, qui sont constituées, par exemple, d'images avec une description textuelle, ou de vidéos et leur sous-titre. L'exploitation de la relation entre ces différentes modalités présente un grand potentiel. Les différentes modalités d'un document peuvent être vues comme une forme de supervision mutuelle, bien que bruitée.

2. **Modélisation d'une sortie structurée** Un grand nombre de méthodes de classification d'images ne modélisent pas explicitement les dépendances entre les étiquettes de classes, alors que bien souvent, ces étiquettes sont implicitement enchâssées



Park bench — *A bench in a public park*

Mortar — *A bowl-shaped vessel in which substances can be ground and mixed with a pestle*

Carousel — *A large, rotating machine with seats for children to ride or amusement*

Brace — *Elastic straps that hold trousers up (usually used in the plural)*

Building Sights; Citylife; No Visual Season; Outdoor; Sky; Night; Architecture; Street; Church; Visual Arts; Natural; Technical; Cute.

Summer; Outdoor; Plants; Trees; Day; Sunny; Neutral Illumination; Partly Blurred; Small Group; Vehicle; Car; Teenager; Adult; Old person.

No Visual Season; No Visual Place; No Visual Time; Neutral Illumination; No Blur; No Persons; Food; Painting; Artificial; Natural; Cute.

Family Friends; Citylife; Sports; No Visual Season; Outdoor; Day; Sunny; Big Group; Vehicle; Overall Quality; Visual Arts; Cute; Skateboard; Teenager.

Figure 1.3 *Illustration de la diversité des concepts utilisés en classification d'images, à l'aide d'images extraites de: (*haut*) la base ImageNet ILSVRC'10 (Berg et al., 2010), présentée avec les noms de classes pertinentes, et leur description, parmi un ensemble de 1,000 classes différentes, et (*bas*) la base ImageClef VCDT'10 (Nowak and Huiskes, 2010) ainsi que certains des concepts pertinents, parmi un ensemble de 93 étiquettes.*

dans une structure. Par exemple, les noms d'animaux pourraient être représentés par la classification classique (ou classification linnéenne), ou la présence de certains objets pourrait promouvoir la présence d'une autre classe avec laquelle ils sont positivement corrélés.

3. **Utilisation d'interactions avec l'utilisateur** L'interaction avec un utilisateur au moment du test offre une opportunité intéressante d'améliorer la performance des algorithmes de classification. La quantité d'interaction manuelle permet un compromis intéressant entre la précision des prédictions, et la quantité de documents qui sont annotés pendant une période de temps donnée.

4. **Classification de classes jamais observées** Les bases de donnée réalistes sont souvent indéterminées, avec des images de classes jusqu'alors jamais vues ajoutées à la base. Le défi à relever est la création de systèmes pour la classification capables de classifier des images pour des classes jamais vues pendant l'apprentissage. Nous considérons deux stratégies populaires, qui permettent la classification de classes jamais observées: la classification basée sur les attributs, et la classification basée sur la distance.

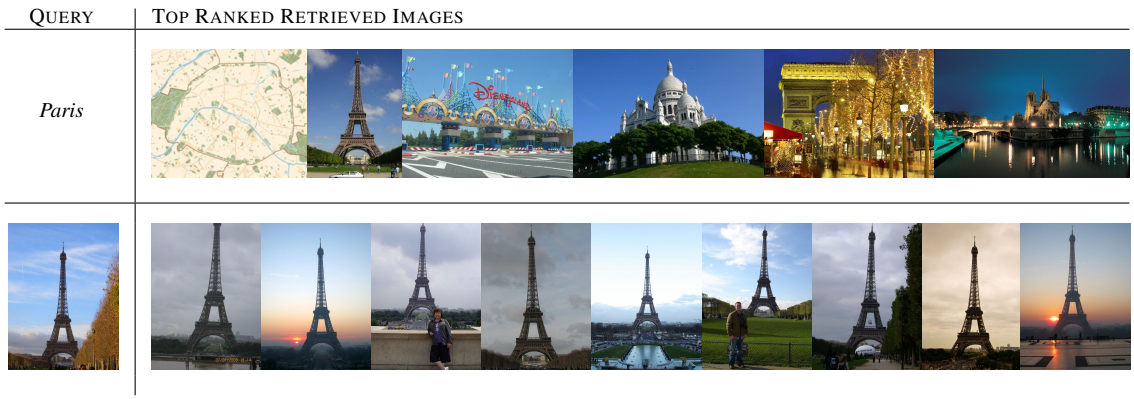


Figure 1.4 *Exemple de systèmes de recherche, (*haut*), montrant les images les mieux classées par l'outil de recherche d'images de Googles pour la requête «*Paris*», et (*bas*) montrant les images les mieux classées pour un système similaire BigImbaz*[1]*, pour l'image requête montrées sur la gauche.*

5. **Passage à l'échelle de grandes bases de données** Ces 5 dernières années, la définition d'une base à grande échelle a changé, passant de bases contenant un peu plus de dix mille images, à des bases contenant plus d'un million d'images. Cette tendance reflète un besoin pour des méthodes d'interprétation d'images efficaces, qui peuvent gérer les bases à grande échelle que l'on croise de nos jours. Idéalement, nous voudrions que nos algorithmes d'apprentissage soient efficaces selon plusieurs critères: le coût de calcul à l'apprentissage et au test, les besoins en stockage mémoire pour les représentation d'images, le nombre d'images d'apprentissage nécessaires à la généralisation, l'effort d'intervention manuelle nécessaire, par exemple pour l'annotation d'images d'apprentissage.

1.4 Plan de la thèse

Dans ce qui suit, nous décrivons le plan et les principaux contenus des chapitres présentés dans ce manuscrit.

Chapitre 2 Dans ce chapitre, nous décrivons de manière détaillée des représentations d'images utilisées pour la classification et la recherche d'images. Nous nous intéresserons plus particulièrement à la représentation par sac-de-mots (Csurka et al., 2004), et au vecteur de Fisher (Perronnin and Dance, 2007; Perronnin et al., 2010).

Nous donnons aussi un aperçu des techniques de classification paramétriques, comme la régression logistique, les séparateurs à vaste marge (SVM) (Vapnik, 1995), mais aussi

[1]Demo disponible http://bigimbaz.inrialpes.fr/

des techniques de classification non-paramètriques, comme les k plus proches voisins (k-PPV), ou le modèle TagProp (Guillaumin et al., 2009). Bien que ce chapitre ne contienne aucune des contributions majeures de la thèse, il définit les concepts nécessaires à la compréhension des chapitres qui suivront.

Chapitre 3 Dans ce chapitre, nous considérons la recherche d'image, et la classification multi-étiquette d'image, utilisant des basées de données multimodales. Plus précisément dans ce cas, nous considérons des images accompagnées de mots-clés et/ou de légendes.

Dans les systèmes de recherche, une approche fructueuse pour l'exploitation de la mutlimodalité est l'utilisation de modèle de retour de pertinence entre média «*transmedia relevance feedback*» qui, pour une image requête donnée, utilisent les descriptions textuelles des plus proches voisins visuels de l'image requête, pour étendre la requête texte. Nous introduisons une paramétrisation pour apprendre les paramètres de ces modèles de retour de pertinence pour la recherche d'images.

Nous avons également étendu le modèle TagProp (Guillaumin et al., 2009) afin d'exploiter ces distances de retour de pertinence pour la classification d'images multi-étiquettes. Nous montrons expérimentalement que l'utilisation de ces paramètres pour des modèles de retour de pertinence sont meilleurs que les modèles actuels de l'état de l'art sur les bases utilisées pour ces deux tâches. Le travail de ce chapitre a été publié en parti dans (Mensink et al., 2010).

Chapitre 4 Dans ce chapitre, nous considérons les tâches d'annotation d'images, et de classification d'images basée sur les attributs. Pour ces tâches, nous introduisons une série de modèles structurés qui exploitent la corrélation entre les étiquettes de classe, et qui utilisent les interactions avec l'utilisateur. Dans ce scénario d'annotation interactive, un utilisateur doit donner la valeur d'un certain nombre d'étiquettes au moment du test. Nous pensons qu'un tel scénario propose un compromis intéressant entre la qualité de la prédiction, et les efforts d'annotation manuelle.

Nous avons utilisé des modèles graphiques à base de structure arborescente généralisée afin de modéliser les dépendances entre les étiquettes d'image et les attributs. Chaque nœud du modèle graphique représente une ou un petit nombre d'étiquettes, choisies parmi le vocabulaire d'annotation.

Alors que ces modèles prennent en compte un grand nombre d'interactions entre étiquettes, leur estimation est tout de même possible.

Nous avons appliqué ces modèles à base de structures arborescentes aux tâches d'annotation d'image, et à l'apprentissage sans exemples «*zero-shot learning*». Pour cette dernière, nous avons utilisé des paradigmes de classification basés sur les attributs. Dans ce cas, les prédictions d'attributs sont combinées avec des associations entre les attributs et les classes, et appliquées à la classification des images de test, pour des classes jamais vues. Des modèles structurés sont appliqués au niveau des attributes. Les modèles structurés

que nous proposons dans ce manuscrit obtiennent de meilleurs résultats que les prédicateurs d'étiquettes ou d'attributs indépendants. Cependant, l'avantage majeur de ces modèles est leur capacité à bénéficier des interactions avec les utilisateurs, par rapport à des classifieurs indépendants.

Nous considérons aussi l'apprentissage de modèles spécifiques aux mesures de rangs utilisées pour l'évaluation. Malheureusement, l'optimisation de plusieurs mesures de performances basées sur le rang et intimement liée aux problèmes d'association quadratique qui sont NP-complets.

Ceci est valable aussi pour les modèles basés sur des arbres, lorsqu'ils sont optimisés pour des mesures de rang. Nous avons identifié une sous-classe de ces problèmes qui peut être résolue en complexité polynomiale, basée sur des modèles structurés c-étoile qui permettent la modélisation d'un grand nombre d'interactions binaires entre les étiquettes. Lorsqu'ils sont combinés avec des fonctions de score spécifiquement conçues pour la mesure précision-à-k, ces modèles peuvent être résolus efficacement, et produisent de meilleurs résultats que les modèles indépendants. Néanmoins, nous observons que les modèles de mélange d'arbres, qui encodent plus d'interactions binaires, mais sont optimisés pour la classification, sont meilleurs que le modèle c-étoile.

Ce travail a été publié dans (Mensink et al., 2011a, 2012b, 2011b).

Chapitre 5 Dans ce chapitre, nous considérons des classifiers par k plus proches voisins (k-PPV), et par plus proche moyenne (PPM), pour la classification d'images à grande échelle. Alors que ces méthodes permettent la généralisation à de nouvelles classes, elles ont tendance à obtenir des résultats peut satisfaisants lorsqu'une distance Euclidienne est utilisée. C'est pourquoi nous proposons d'explorer l'apprentissage de métriques partagées entre les classes.

Les méthodes proposées sont efficaces, et permettent un apprentissage sur la base «*ImageNet Large Scale Visual Recognition Challenge 2010*» qui contient plus de 1,2 millions d'images de 1 000 classes, tout en utilisant une représentation d'image de soixante-quatre mille dimensions. De façon surprenante, nous avons observé que le classifier PPM, avec une métrique apprise, obtient des performances similaires aux SVMs, et surpasse le classifier k-PPV, plus flexible. Puisque le classifier PPM est un classifier linéaire, il permet une classification efficace.

Pour la méthode PPM, nous montrons la généralisation à une base de données contenant plus de 10 millions de classes, tout en utilisant une métrique apprise sur la base ILSVRC'10. De façon surprenante, la méthode PPM obtient des résultats compétitifs avec des classifieurs appris de façon spécifiques pour ces 10 mille classes, alors qu'il se contente de calculer la moyenne de ces classes. Ces bonnes performances, combinées avec la possibilité de classifier des classes qui n'ont pas été vues pendant l'apprentissage fait du classifier PPM une alternative intéressante au classifier SVM.

Ce travail a été publié dans (Mensink et al., 2012a, Submitted2012).

1.5 Perspectives

Ils existent de très nombreuses façon d'étendre les méthodes décrites dans cette thèse. Ici, nous détaillons trois idées spécifiques à poursuivre, directement inspirées des idées présentées dans ce manuscrit.

Apprentissages actif et interactif Les systèmes de reconnaissance d'images actuels sont toujours loin derrière les performances et la robustesse du système visuel humain, et n'ont pas la même capacité à utiliser le contexte entre les objets. Dans cette thèse, nous avons utilisé un système d'annotation interactif des images, qui a obtenu des performances significativement plus élevées que des systèmes de prédiction totalement automatiques, même si quelques étiquettes ont été fournies par un utilisateur. Ce scénario interactif est très proche du concept d'apprentissage actif. Ces deux scénarios visent à étiqueter des images avec une quantité limitée d'annotation manuelle.

Nous pourrions mettre un place un système de classification dont le but serait d'obtenir rapidement l'étiquetage précis d'une image, au coût le plus bas possible. Le vocabulaire de ce système serait déterminé par l'intérêt de l'utilisateur, en d'autres termes, l'utilisateur décide quelles étiquettes devraient être utilisées dans le processus d'annotation.

Nous proposons les questions de recherche suivantes:

- comment apprendre une structure entre les étiquettes, alors que le vocabulaire est en constante évolution;
- comment détecter que de nouvelles classes sont nécessaires;
- comment déterminer un score de confiance pour chaque étiquette fournie par l'utilisateur pour permettre de gérer des étiquettes erronées, et
- comment assurer que l'utilisateur produise des étiquettes de grande qualité.

Prédiction structurée, en utilisant l'apprentissage local Dans le chapitre 4, nous avons introduit une série de modèles structurés paramétriques, et dans les chapitres 2, 3 et 5, nous avons montré le succès des approches basées sur l'apprentissage local, comme k-PPV et TagProp. Il serait intéressant de développer une méthode combinée, qui permette l'apprentissage local pour la prédiction structurée. Supposons que nous sommes intéressés par la classification d'images multi-étiquette, c'est à dire prédire conjointement la valeur d'un ensemble d'étiquettes. Nous pourrions appliquer des méthodes d'apprentissage local, par l'affectation d'un point à chaque image de la base, basée sur des distances visuelles, et propager ces annotations aux images de test.

Pour permettre une prédiction structurée, nous pourrions propager les marginales binaires des plus proches voisins vers les images de test, c'est à dire utiliser un compte pondéré de chaque configuration possible pour les paires de labels. Nous pourrions modéliser les étiquettes dans une structure arborescente, et obtenir des marginales pour des paires d'étiquettes, et pour des étiquettes seules, à partir des voisins des images de test.

Parmi les directions de recherche possibles, citons:

- comment définir une structure arborescente et l'adapter au voisinage local d'une image de test
- comment ajouter de la connaissance a priori dans le modèle; et
- comment modéliser un compromis entre des marginales binaires, apprises localement et globalement

Représentation des classes Dans beaucoup de systèmes de classification, on suppose qu'une classe peut être représentée par un simple vecteur de poids (par exemple pour les SVMs), ou par un simple vecteur moyen (par exemple pour NCM). Cependant, étant donnée la grande variabilité intra-classe, combinée avec d'autres difficultés inhérentes de la perception visuelle, comme les variations de points de vues, ou les scènes complexes, il est en fait surprenant qu'une représentation unique pour chaque classe fonctionne si bien.

Des représentations de classe plus riches pourraient être utilisées de différentes façons, par exemple

- Des moyennes par classe, utilisées pour le classifier NCMC (voir chapitre 5), auraient pu être obtenues en alternant un regroupement par les k moyennes, et un apprentissage de métrique, pour obtenir de meilleurs représentants par classe.
- Des représentants par classes plus discriminants pourraient être obtenus à l'aide d'une méthode de regroupement tenant compte de toutes les classes simultanément, et non indépendemment.
- Dans le cas de classifieurs SVM, nous pourrions utiliser des modèles de SVM à variables latentes pour la classification, ou les variables latentes seraient utilisées pour associer chaque image à une «sous-classe».
- Dans le contexte de la classification multi-étiquette, nous pourrions apprendre des classifieurs conjoints, de la même façon que les «phrases visuelles» sont utilisées pour la détection d'objet. Par exemple, un classifier conjoint pour «voiture-vélo» un-contre-tous serait d'un grand intérêt pour le cas où les deux objets apparaissent ensembles, par rapport à la combinaison de deux classifiers voiture et vélo indépendants.

Bibliography

A. Berg, J. Deng, and F.-F. Li. The ImageNet large scale visual recognition challenge 2010. http://www.image-net.org/challenges/LSVRC/2010, 2010.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the International Conference on Computer Vision*, 2009.

T. Mensink, J. Verbeek, and G. Csurka. Trans media relevance feedback for image autoannotation. In *Proceedings of the British Machine Vision Conference*, 2010.

T. Mensink, J. Verbeek, and G. Csurka. Learning structured prediction models for interactive image labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011a.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *Proceedings of the European Conference on Computer Vision*, 2012a.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Large scale metric learning for distance-based image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Submitted2012.

T. Mensink, J. Verbeek, and T. Caetano. Learning to rank and quadratic assignment. In *NIPS Workshop on Discrete Optimization in Machine Learning*, Granada, Spain, December 2011b.

T. Mensink, J. Verbeek, and G. Csurka. Tree-structured crf models for interactive image labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012b.

S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at ImageCLEF 2010. In *Working Notes of CLEF*, 2010.

V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, 2011.

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision*, 2010.

A. 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, 2000.

R. Szeliski. *Computer Vision: Algorithms and Applications*. Spinger-Verlag, 2011.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.