ÉCOLE CENTRALELYON

UNIVERSITÉ DE LYON

Habilitation à Diriger des Recherches

Ecole Centrale de Lyon

École Doctorale InfoMaths (Université de Lyon)

Spécialité : Informatique

Contributions à la compréhension automatique de données visuelles

Par : **Emmanuel DELLANDRÉA**

Composition du jury :

Rapporteurs :
Pr. Christine Fernandez-Maloigne, Université de Poitiers
Pr. Su Ruan, Université de Rouen
Dr. Benoit Huet, EURECOM

Examinateurs :
Pr. Martha Larson, Université de Radboud, Pays-Bas
Pr. Jenny Benois-Pineau, Université de Bordeaux
Dr. Georges Quénot, Laboratoire d'Informatique de Grenoble
Pr. Mohand Saïd Hacid, Université Lyon 1
Pr. Liming Chen, Ecole Centrale de Lyon

Date de soutenance : 12 juin 2020

Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)
UMR 5205

Table des matières

Introduction

Ce mémoire synthétise les principales contributions de mes activités de recherche depuis mon recrutement comme Maître de Conférences en 2004 à l'Ecole Centrale de Lyon. La problématique générale à laquelle ces travaux s'attachent est la compréhension automatique de scènes visuelles. L'objectif est ainsi d'élaborer des algorithmes et méthodes permettant d'identifier des objets et des concepts d'intérêt dans les images et vidéos. Cela relève donc des domaines de la vision par ordinateur et de l'apprentissage automatique.

Contexte

Depuis plusieurs années, nous assistons à une croissance exponentielle de la quantité de données visuelles (images et vidéos) disponibles à tout un chacun à partir d'archives en ligne, de sites sociaux de partage ou encore de collections professionnelles et personnelles. Face à ce phénomène, il est apparu nécessaire de développer des outils efficaces pour permettre l'organisation, la recherche, la classification et l'interprétation de ces collections de données. Ceci a provoqué et continue de provoquer une émulation extrêmement importante dans les communautés de vision par ordinateur et d'apprentissage automatique comme en témoignent notamment les nombreuses compétitions dont l'objectif est d'extraire automatiquement de l'information sémantique sur le contenu des images directement à partir des valeurs des pixels de ces images, telles que Pascal VOC [EEVG+15], TRECVID [ABC+18] , ImageCLEF [MCDC12], ImageNet Large Scale Visual Recognition Challenge [DDS+09] ou encore COCO Challenge [LMB+14]. Ainsi, de nouvelles problématiques et d'importants verrous scientifiques sont apparus. La principale difficulté est connue comme le «fossé sémantique» caractérisant le fait que des concepts de haut-niveau sémantique tels que «chien», «personne», «voiture», «sentiment de stress» doivent être identifiés par l'ordinateur à partir des données bas-niveau que sont les pixels de l'image. Par ailleurs le nombre d'applications de ces techniques ne cessent de croître notamment dans des domaines actuellement très porteurs et stratégiques tels que la recherche «intelligente» d'information visuelle, la médecine, la conduite autonome ou encore la robotique.

Tendances et évolutions méthodologiques

Durant la période couverte par ce mémoire, des évolutions méthodologiques majeures se sont produites dans le domaine de la compréhension automatique de scènes visuelles.

Dans les années 2000, la tendance consistait à procéder en deux étapes : une première phase d'extraction de descripteurs à partir des données visuelles puis une phase de classification s'appuyant sur ces descripteurs pour prédire les concepts d'intérêt associés aux données. Ces descripteurs devaient être choisis de manière à capturer les propriétés visuelles des éléments à reconnaître, être robustes aux occlusions partielles, de posséder des propriétés d'invariance, notamment au point de vue et à l'illumination. L'approche dominante était de calculer des descripteurs locaux, puis de les agréger par un apprentissage non supervisé de manière à obtenir des descripteurs globaux au niveau de l'image, et enfin d'appliquer un classifieur par apprentissage supervisé pour réaliser la classification. Parmi les descripteurs locaux les plus efficaces et populaires figuraient les descripteurs HOG [DT05] et SIFT [Low04], agrégés par histogrammes de sacs de mots visuels [CDF+04] ou par vecteurs Fisher [PD07]. Des méthodes à noyaux telles que les machines à vecteurs supports (SVM) [CV95] étaient alors appliquées pour la classification.

L'année 2012 a marqué un tournant fondamental dans le domaine, avec le succès de Krizhevsky et al. au challenge ImageNet Large Scale Visual Recognition Challenge [KSH12] proposant une approche reposant sur un réseau de neurones convolutif, inspiré de [CBD+90]. Le gain de performances par rapport aux méthodes reposant sur le paradigme précédent fut tel quel cela ouvrit la voie au deep learning et à ce qui est dorénavant communément appelé "Intelligence Artificielle". La grande force de ce nouveau paradigme est de permettre un apprentissage cohérent basé sur les données et au sein du même réseau de neurones, à la fois des descripteurs par les couches de convolution, et du classifieur par les couches entièrement connectées, d'une manière bout à bout en prenant en entrée du réseau l'image elle-même, et en fournissant en sortie la catégorie ou l'objet dans l'image. Bien que prêtes à éclore depuis les années 90 sous l'impulsion de Yann LeCun [CBD+90], il aura fallu attendre l'année 2012 et la conjonction de deux phénomènes révolutionnaires dans le domaine pour voir apparaître le succès de cette famille de méthodes, à savoir la convergence de la puissance de calculs parallèles offerte par les cartes graphiques, et la quantité de données proposée par la communauté, et en particulier la base ImageNet avec ses 14 millions d'images annotées selon environ 20 mille catégories.

Il y a fort à parier que ces méthodes extrêmement performantes et prometteuses

ont encore de beaux jours devant elles. Elles offrent notamment de nouveaux champs d'investigation en posant de nouvelles problématiques, en particulier l'interprétabilité des prédictions fournies par ces réseaux de type "boîte noire", et la nécessité d'imaginer des solutions pour contrecarrer le fait que pour leur apprentissage, un nombre très important de données est a priori nécessaire.

Contributions et organisation du mémoire

Mes activités de recherche et principales contributions dans ce domaine de la compréhension automatique de scènes visuelles portent sur la classification d'images, la détection d'objets dans les images ainsi que l'application de l'analyse visuelle à la prédiction de l'impact émotionnel des vidéos et à la robotique. Ces travaux sont le résultat de collaborations avec plusieurs doctorants et un résumé est donné dans les chapitres suivants.

Dans le chapitre 1, nous présentons nos travaux dédiés à la classification d'images.

Le chapitre 2 est consacré à nos contributions pour la problématique de la détection des objets dans les images.

Dans le chapitre 3, nous détaillons nos travaux portant sur la prédiction de l'impact émotionnel des vidéos.

Nos contributions dans le domaine de l'analyse visuelle pour la robotique sont présentées dans le chapitre 4.

Enfin, le chapitre 5 conclut ce mémoire par un résumé de nos contributions et présente plusieurs directions pour nos prochains travaux.

Chapitre 1

Classification d'images

1.1 Introduction

La classification d'images consiste à associer automatiquement à l'image des étiquettes indiquant les concepts de haut-niveau sémantique identifiés dans l'image, tels que des scènes (intérieur, extérieur, paysage, ...), des objets (voiture, animal, personne, ...), des événements (voyage, travail, ...), ou encore des émotions (joie, mélancolie, ...). Typiquement, un système de classification d'images s'appuie sur l'extraction de descripteurs pour caractériser les données, la sélection des descripteurs les plus pertinents, puis l'application d'un modèle de prédiction préalablement appris, suivie éventuellement d'une phase de fusion des décisions fournies pas plusieurs modèles de prédiction. Ainsi, le système permet d'obtenir en sortie des scores indiquant la probabilité pour les concepts cibles d'être présents dans l'image d'entrée.

Nous avons proposé plusieurs contributions dans ce domaine, touchant aux différentes étapes de la chaîne d'analyse des images et brièvement résumées ci-dessous. Ces travaux ont été réalisés avec les doctorants Huanzhang Fu et Ningning Liu, notamment dans le cadre des projets ANR Omnia et VideoSense et ont donné lieu aux publications [R5, R10, L1, C12, C13, C15, C24, C26, W4] (numérotation correspondant aux publications listées dans le rapport d'activité en fin de document).

1.2 Descripteurs basés sur une représentation parcimonieuse des images

L'objectif d'une représentation parcimonieuse est d'obtenir une représentation fidèle d'un signal pouvant être considéré comme une combinaison linéaire d'atomes constituant un dictionnaire de dimension très supérieure à celle du signal lui-même [MZ93]. Cette décomposition va introduire dans la nouvelle représentation du signal un grand nombre de valeurs nulles, d'où la parcimonie. Elle a été originellement

proposée dans le domaine du traitement du signal comme un outil puissant pour acquérir, représenter et compresser des signaux de grande dimension, avec notamment des applications en séparation de sources [GR13]. Des études ont également montré que ces principes s'appliqueraient aux neurones du cortex visuel qui utiliseraient un codage parcimonieux pour représenter efficacement des scènes naturelles [OF96, OF97]. Ainsi, ces techniques ont commencé à impacter le domaine de la vision par ordinateur, avec des travaux notamment sur le débruitage d'images [LCHR19] la reconnaissance de visages [WYG+09], la super-résolution d'images [JWHY08], l'extraction de caractéristiques locales d'images [MBP+08], la segmentation de mouvement [RTVY08] ou encore la modélisation de l'arrière-plan [DH08]. Ces intéressantes propriétés et facultés nous ont conduit à proposer une adaptation de ces principes au problème de la classification d'images.

Dans ce cadre, nous avons développé une représentation parcimonieuse reconstructive et discriminative des images, intégrant un terme discriminant tel que la mesure discriminative de Fisher [Bis06] ou encore la sortie d'un classifieur, dans la fonction objectif de la représentation parcimonieuse afin de permettre l'apprentissage d'un dictionnaire reconstructif et discriminatif.

Rappelons tout d'abord les principes d'une représentation parcimonieuse.

Soit un signal $y \in \mathbb{R}^n$ qui sera représenté par une combinaison linéaire d'éléments de base d'un dictionnaire $D \in \mathbb{R}^{n \times K}$ composé d'atomes en colonne $\{d_j\}_{j=1}^K$. Une représentation du signal y basé sur ce dictionnaire D est tout vecteur $x \in \mathbb{R}^K$ satisfaisant :

$$y = Dx \tag{1.1}$$

Dans le cas où $n < K$, le dictionnaire D est dit sur-complet et cette équation a alors de nombreuses solutions possibles. Généralement, la solution correspondant au minimum de la norme l^2 est alors choisie :

$$\min_x(||x]]_2) \text{ subject to } Dx = y \tag{1.2}$$

où $||x||_2$ est la norme l^2 de x. Ce problème peut être aisément résolu et son unique solution est donnée par :

$$x = D^+ y = D^T(DD^T)^{-1}y \tag{1.3}$$

où D^+ est la matrice pseudoinverse de D. Cependant, cette solution n'est généralement pas parcimonieuse et contient un nombre important d'éléments non nuls

correspondant aux atomes du dictionnaire et ainsi ne correspond pas à notre objectif. En effet, une solution parcimonieuse serait plus favorable, avec une combinaison linéaire d'un nombre très restreint d'atomes pour approximer le signal y. Ce problème peut être formulé de la manière suivante :

$$\min_x(||x||_0) \text{ subject to } Dx = y \tag{1.4}$$

où $||x||_0$ est la norme l^0 de x qui est égale au nombre d'éléments différents de 0 dans le vecteur x.

La résolution de l'équation (1.4) est un problème NP difficile. Néanmoins, de nombreuses techniques d'approximation ont été proposées telles que Matching Pursuit (MP) [MZ93] qui est un algorithme glouton consistant à sélectionner un à un les atomes afin de minimiser l'erreur résiduelle, ou encore Orthogonal Matching Pursuit (OMP) [PRK93] impliquant le calcul des produits scalaires entre le signal et les atomes du dictionnaire.

Un autre aspect crucial pour appliquer un modèle de représentations parcimonieuses avec succès sur un signal ou une image, est la construction du dictionnaire D. Une première stratégie consiste à utiliser un dictionnaire prédéfini qui ne sera pas modifié durant la résolution du problème. De tels dictionnaires basés sur des transformées telles que ridgelet, curvelet ou contourlet sont largement utilisées en traitement du signal [SED05, OSL00, EA06]. Une autre stratégie consiste à utiliser un dictionnaire constitué des signaux/images d'apprentissage, ce qui a également donné de bons résultats [FZD+09, WYG+09]. Cependant, dans certaines situations, il n'existe pas de bases appropriées efficaces pour représenter les signaux. Ainsi, une troisième stratégie a été proposée afin de permettre l'apprentissage d'un dictionnaire spécifique à une tâche donnée, composé initialement d'exemples qui seront mis à jour afin de décrire au mieux le contenu des images. Parmi les méthodes les plus populaires dans ce contexte figurent Method of Optimal Directions (MOD) [EAH99] et K-SVD [AEB06]. Toutes deux sont des méthodes itératives, contenant une phase de codage parcimonieux afin d'identifier les coefficients x appropriés pour un signal y basé sur le dictionnaire courant, suivie d'une phase de mise à jour du dictionnaire utilisant les coefficients obtenus précédemment pour mieux s'adapter aux données.

S'appuyant sur ces résultats, l'approche que nous avons développée, Reconstructive and Discriminative Sparse Representation for Visual Object Categorization (RDSR_VOC) est présentée ci-dessous.

Soit un ensemble N de signaux d'entraînement $\{y_i\}_{i=1}^N$ appartenant à M catégories. $Y = [y_1, y_2, \ldots, y_N]$ est la matrice contenant les signaux en colonne, et

$X = [x_1, x_2, x_N]$ sont les coefficient parcimonieux associés basés sur le dictionnaire D. De plus, supposons que N_i signaux sont de la catégorie M_i, pour $1 \leq i \leq M$.

La fonction objectif de la représentations reconstructive parcimonieuse standard est donnée par :

$$\min_{D,X}\{||Y - DX||_F^2\} \text{ subject to } ||x_i||_0 \leq L \ \forall i \tag{1.5}$$

où $||A||_F$ est la norme de Frobenius définie par $||A||_F = \sqrt{\sum_{i,j} a_{i,j}^2}$ et L est nombre positif contrôlant le niveau de parcimonie.

Lorsque la contrainte de parcimonie est intégrée à la fonction, cela donne :

$$\min_{D,X,\Lambda} \{\lambda_1 ||Y - DX||_F^2 + \lambda_2 \sum_{i=1}^{N} ||x_i||_0\} \tag{1.6}$$

ou encore

$$\min_{D,X,\Lambda} \{\lambda_1 \sum_{i=1}^{N} ||y_i - Dx_i||_2^2 + \lambda_2 \sum_{i=1}^{N} ||x_i||_0\} \tag{1.7}$$

où $\Lambda = \{\lambda_1, \lambda_2\}$ est un ensemble de paramètres de régularisation ajustant le poids donné à l'erreur de reconstruction par rapport à la parcimonie.

L'objectif principal de notre approche est d'apprendre un dictionnaire reconstructif et discriminatif permettant d'augmenter le pouvoir discriminatif de la représentation parcimonieuse du signal basée sur le dictionnaire, tout en conservant une faible erreur de reconstruction, c'est à dire que le signal reconstruit à partir des coefficients parcimonieux doit être aussi proche que possible du signal original. Ainsi, inspiré par [HA07], le terme discriminant de Fisher est introduit dans la fonction objectif.

Supposons que S_W est la covariance intra-classe :

$$S_W = \sum_{i=1}^{M} S_i \tag{1.8}$$

où

$$S_i = \sum_{x_j \in M_i} (x_j - m_i)(x_j - m_i)^T \tag{1.9}$$

avec

$$m_i = \frac{1}{N_i} \sum_{x_j \in M_i} x_j \tag{1.10}$$

m_i et N_i sont respectivement la moyenne des signaux et le nombre de signaux appartenant à la catégorie M_i.

Soit S_B la covariance inter-classe :

$$S_B = \sum_{i=1}^{M} N_i (m_i - m)(m_i - m)^T \tag{1.11}$$

où m est la moyenne de l'ensemble des signaux

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1.12}$$

Ainsi, le terme discriminatif de Fisher s'exprime selon l'équation (1.13) :

$$F(X) = \frac{||S_B||_2^2}{||S_W||_2^2} = \frac{|| \sum_{i=1}^{M} N_i (m_i - m)(m_i - m)^T ||_2^2}{|| \sum_{i=1}^{M} \sum_{x_j \in M_i} (x_j - m_i)(x_j - m_i)^T ||_2^2} \tag{1.13}$$

Le terme discriminatif de Fisher est maximisé lorsque la distance entre les classes est maximisée alors que celles à l'intérieur des classes sont minimisées, ce qui est censé améliorer la classification.

L'incorporation du terme de Fisher dans l'équation (1.7) donne alors :

$$\min_{D,X,\Lambda} \{ \lambda_1 \sum_{i=1}^{N} ||y_i - D x_i||_2^2 + \lambda_2 \sum_{i=1}^{N} ||x_i||_0 - \lambda_3 F(X) \} \tag{1.14}$$

où $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ est, comme pour l'équation (1.7), un ensemble de paramètres de régularisation ajustant le poids donné à l'erreur de reconstruction, à la parcimonie et au pouvoir discriminatif de Fisher.

Le dictionnaire reconstructif et discriminatif peut alors être appris en résolvant le problème de minimisation formulé dans l'équation (1.14). Ainsi, la représentation parcimonieuse qui gagne en pouvoir discriminatif tout en restant fidèle au signal d'origine peut également être obtenue par un codage parcimonieux basé sur le dictionnaire appris.

Comme indiqué précédemment, l'apprentissage d'un dictionnaire se fait généralement par des méthodes telles que MP et OMP. Cependant, notre fonction reconstructive et discriminative, ne met pas en jeu un seul signal, mais tous les signaux d'apprentissage. Ainsi, ces méthodes ne peuvent pas être appliquées directement. Nous avons donc proposé un algorithme Sequential Forward Sparse Coding (SFSC) pour réaliser cette tâche.

Soit G la fonction à minimiser :

$$G = \lambda_1 \sum_{i=1}^{N} ||y_i - Dx_i||_2^2 + \lambda_2 \sum_{i=1}^{N} ||x_i||_0 - \lambda_3 F(X) \tag{1.15}$$

La première étape de SFSC consiste à sélectionner l'atome du dictionnaire D avec la plus petite valeur de G en considérant que seul cet atome a été utilisé pour la décomposition parcimonieuse afin d'obtenir les coefficients parcimonieux de tous les signaux $\{x_i\}_{i=1}^{N}$. En effet, si le sous-ensemble Γ des indices des atomes utilisés pour la décomposition parcimonieuse est connu, les coefficients parcimonieux peuvent alors être obtenus par :

$$X = D_\Gamma^+ Y \tag{1.16}$$

où D_Γ est le dictionnaire réduit uniquement composé des atomes dont les indices sont dans Γ et D_Γ^+ est la pseudo-inverse de la matrice D_Γ. A chaque étape suivante, un nouvel atome sera sélectionné parmi ceux restants, correspondant à la valeur minimale de G basé sur le sous-ensemble des atomes formés par la combinaison des atomes sélectionnés précédemment complété de ce nouveau, jusqu'à ce qu'un critère d'arrêt soit atteint. Il peut s'agir du nombre attendu d'atomes à utiliser pour la décomposition parcimonieuse, ou encore le moment auquel la valeur de G commence à augmenter. L'algorithme détaillé est donné dans la Figure 1.1.

Concernant la mise à jour du dictionnaire, il est possible d'utiliser la méthode K-SVD [AEB06]. L'algorithme complet ainsi obtenu, RDSR_VOC, est donné dans la Figure 1.2.

Un avantage de RDSR est que d'autres critères discriminatifs peuvent être aisement utilisés en remplaçant $F(X)$ de la fonction objectif sans avoir à modifier le reste de l'algorithme. Cela peut par exemple être le taux de classification d'un classifieur, ce que nous avons évalué dans nos expérimentations. Celles-ci ont été conduites sur le jeu de données SIMPLIcity [WJW01] contenant un ensemble de 1000 images de 10 catégories (100 images par catégorie) : African & village(C1), Beach (C2), Building (C3), Bus (C4), Dinosaur (C5), Elephant (C6), Flower (C7), Horse (C8), Mountain & glacier (C9) et Food (C10). Quelques exemples sont donnés dans la Figure 1.3.

Un total de 2246 descripteurs ont été extraits pour caractériser les propriétés locales et globales de chaque image en terme de couleur, texture et forme.

A des fins de comparaison, trois types de modèles ont été évalués. Le premier consiste en un classifieur SVM [CV95] prenant en entrée les descripteurs image, noté SVM dans la suite. Le deuxième modèle, RSR_VOC, s'appuie sur une représenta-

SFSC algorithm

- Task: Given the dictionary $D \in \mathbb{R}^{n \times K}$, the regularization parameter set Λ and the set of signal $Y = [y_1, y_2, ..., y_N]$ to be represented by a linear combination of atoms from D, find the corresponding coefficients $X = [x_1, x_2, ..., x_N]$ that minimize G

$$G = \lambda_1 \sum_{i=1}^{N} ||y_i - Dx_i||_2^2 + \lambda_2 \sum_{i=1}^{N} ||x_i||_0 - \lambda_3 F(X).$$

- Initialization: Set the initial index set $\Gamma^0 = \emptyset$ and the indicator of iteration $t = 1$.
- Repeat until stopping rule:
  - For each $i \notin \Gamma^{t-1}$ and $i \in \{1, 2, ..., K\}$, let $\Psi = \Gamma^{t-1} \cup i$. Then calculate the sparse coefficients $X = D_\Psi^+ Y$ as well as the value of G_i based on X. D_Ψ represents the reduced dictionary composed by the columns in D whose indices are in Ψ
  - $i_{min} = \arg_i \min(G_i)$
  - $\Gamma^t = \Gamma^{t-1} \cup i_{min}$
  - $t = t + 1$
- Calculate the sparse coefficients $X = D_{\Gamma^t}^+ Y$.

FIGURE 1.1 – Algorithme SFSC.

tion parcimonieuse des images uniquement reconstructive. Dans ce cas, le dictionnaire contient les vecteurs de descripteurs des images de l'ensemble d'apprentissage (chaque colonne du dictionnaire est un vecteur de descripteurs) et la fonction objectif est purement reconstructive (le terme discriminatif est exclu). L'optimisation est alors réalisée par l'algorithme OMP pour obtenir les coefficients parcimonieux. La représentation parcimonieuse des images faite de ces coefficients est alors utilisée pour alimenter des classifieurs SVM pour permettre la classification. Enfin, le troisième modèle s'appuie sur une représentation parcimonieuse reconstructive et discrimitative obtenue par l'algorithme RDSR. Trois critères discriminatifs ont été évalués : la terme discriminatif de Fisher, la sortie d'un classifieur SVM avec un noyau RBF, et la sortie d'un SVM avec un noyau linéaire, respectivement notés Fisher, SVM_RBF et SVM_Linear ci-dessous.

Les résultats sont donnés dans la Figure 1.4. Les taux de classification moyens, obtenus par validation croisée à 4 échantillons, sont indiqués pour les 10 catégories d'images (en lignes) en utilisant différentes approches (en colonnes).

Ces résultats montrent qu'utiliser une représentation parcimonieuse permet d'augmenter significativement les performances de classification. De plus, une représenta-

RDSR_VOC algorithm

1. Extract the feature vector representing the image visual content for all the images: $f_{i,j} \in \mathbb{R}^n, i \in \{1,2,...,M\}, j \in \{1,2,...,N_i\}$, where M is the number of categories and N_i is the number of images for i-th category.
2. Normalize all $f_{i,j}$ to have unit ℓ^2 norm.
3. Learn a reconstructive and discriminative dictionary D of sparse representation based on training images, by iteratively running the following two stages with the purpose of minimizing the objective function G. D is initialized by a subset of training image vectors, chosen randomly.
   - *Sparse Coding* using SFSC.
   - *Dictionary Update* similar to the dictionary update stage of K-SVD $k = 1,2,...,K$ in D^{t-1}, update it by
     - Define the group of signals that use this atom $\omega_k : \{i|1 \leq i \leq N, x_T^k(i) \neq 0\}$ where x_T^k is the k-th row of X.
     - Compute the overall representation error matrix, E_k, by
     $$E_k = Y - \sum_{j \neq k} d_j x_T^j.$$
     - Restrict E_k by choosing only the columns corresponding to ω_k, and obtain E_k^R.
     - Apply SVD decomposition $E_k^R = U\Delta V^T$. Choose the updated dictionary column $\tilde{d}_k$ to be the first column of U. Update the coefficient vector x_R^k to be the first column of V multiplied by $\Delta(1,1)$. Here x_R^k is a reduced version of the row vector x_T^k by discarding of the zero entries.
4. Compute the sparse coefficients of all the images based on the learned dictionary D, including the training images and test images.
5. Use a classifier (SVM for example) to accomplish the classification task, using the obtained sparse coefficients as input.

FIGURE 1.2 – Algorithme RDSR.

tion parcimonieuse reconstructive et discriminative se révèle être plus efficace qu'une représentation purement reconstructive, ce qui conforte l'idée qu'ajouter un terme discriminatif dans la fonction objectif est plus approprié pour une tâche de classification. L'ensemble des expérimentations plus détaillées peuvent être trouvées dans [FZD+09] et [FDC11].

FIGURE 1.3 – Exemples d'images du jeu de données SIMPLIcity. De gauche à droite et de haut en bas, elles appartiennent aux classes suivantes : African & village, Beach, Buildind, Bus, Dinosaur , Elephant , Flower , Horse , Mountain & glacier et Food.

	SVM	RSR_VOC	RDSR_VOC (Fisher)	RDSR_VOC (SVM_RBF)	RDSR_VOC (SVM_Linear)
C1	80%	90%	88%	86%	85%
C2	82%	73%	78%	76%	74%
C3	62%	78%	82%	79%	77%
C4	84%	97%	96%	95%	97%
C5	100%	100%	100%	100%	100%
C6	86%	86%	83%	85%	82%
C7	84%	95%	97%	97%	97%
C8	98%	98%	94%	94%	95%
C9	72%	73%	82%	77%	78%
C10	86%	85%	91%	87%	91%
Average	83.4%	87.5%	89.1%	87.6%	87.6%

FIGURE 1.4 – Taux de classification moyens pour les 10 catégories d'images pour l'ensemble de test du jeu de données SIMPLIcity (en lignes) en utilisant différentes approches (en colonnes).

1.3 Descripteur textuel pour la caractérisation des images

Afin de compléter les informations portées par les descripteurs visuels, nous avons proposé un nouveau descripteur textuel dédié au problème de la classification d'images. En effet, la plupart des photos publiées sur des sites de partage en ligne (Flickr, Facebook, ...) sont accompagnées d'une description textuelle sous la forme de mots-clés ou d'une légende. Ces descriptions constituent une riche source d'infor-

mation sur la sémantique contenue dans les images et il est donc particulièrement intéressant de les considérer dans un système de classification d'images. Cependant, alors qu'Internet regorge d'images accompagnées de description textuelles, celles-ci sont généralement très courtes, en moyenne de moins d'une dizaine de mots. Un exemple est donné dans la Figure 1.5 où une image de paon est associée à des tags d'utilisateurs tels que «bird», «beautiful» ou encore «interestingness».

{0A432C9F-1732-45E6-90F7-A6A7B75FA889}.jpg

Flickr user tags: peacock, bird, beautiful, pretty, feathers, waimea, waimeafalls, explore, animal, interestingness

FIGURE 1.5 – Un exemple d'image issue du site Flickr associée à des tags d'utilisateurs peu nombreux mais incluant des concepts sémantiques tels que «bird», «beautiful», «interestingness», etc.

Lorsqu'il s'agit d'analyser un document textuel, l'approche dominante s'appuie sur une représentation du texte comme un sac de mots (Bag-of-Words) et une description selon un modèle d'espace vectoriel [SWY75] sous forme d'un vecteur de termes dont chaque composant correspond à nombre d'occurences/fréquence d'un mot. Une des approches statistiques les plus classiques dans ce domaine est TF-IDF (Term Frequency-Inversed Document Frequency), permettant d'évaluer l'importance d'un terme dans un document relativement à un corpus. Ce modèle a donné lieu à plusieurs extensions dont l'analyse sémantique latente (LSA) [Dum04], LSA probabiliste [Hof99] et l'allocation de Dirichlet latente [BNJ03]. L'inconvénient majeur de ces méthodes basées sur l'analyse statistique des occurrences de mots est leur difficulté à intégrer les notions sémantiques. En effet, un texte est simplement interprété comme une collection non ordonnée de mots sans prise en compte de la grammaire ni même de l'ordre des mots. De plus, un document textuel est ainsi simplement résumé en un vecteur de fréquence des mots sans prendre en considération les relations sémantiques entre mots. Plusieurs travaux ont tenté d'apporter une so-

lution à ces problèmes, en particulier en utilisant des structures linguistiques [FE08] avec des termes composés tels que «système d'exploitation», des relations binaires telles que sujet-verbe ou encore des représentations de termes distribuées [LSZ04] proposant de caractériser la signification d'un terme à partir de son contexte (les autres termes avec lesquels il apparaît fréquemment). Bien que ces représentations ne se soient pas montrées plus efficaces que l'approche de sac de mots standard pour des tâches de catégorisation et indexation de texte [MB04], elles représentent cependant un pas en avant dans la prise en compte des liaisons entre mots grâce à leur contexte [SC04]. Cependant, ce contexte reste limité au document traité.

Ainsi, nous avons proposé HTC (Histograms of Textual Concepts) afin ce capturer les relations sémantiques entre concepts. HTC s'inspire du modèle de vecteurs conceptuels [SLP02] qui décrit la signification d'un mot par ses atomes, composants, attributs, comportement, idées associées, etc. Par exemple, le concept «pluie», peut être décrit par «eau», «liquide», «précipitations», «mousson», etc.

Ainsi, l'idée générale derrière HTC est de représenter un document textuel comme un histogramme de concepts textuels selon un dictionnaire (ou vocabulaire), pour lequel chaque valeur associée à un concept est l'accumulation de la contribution de chaque mot du texte pour ce concept, en fonction d'une mesure de distance sémantique donnée.

Pour un dictionnaire D et une mesure de similarité sémantique donnée S, HTC peut être simplement extrait à partir de tags d'une image par un processus à trois étapes décrit dans la Figure 1.6.

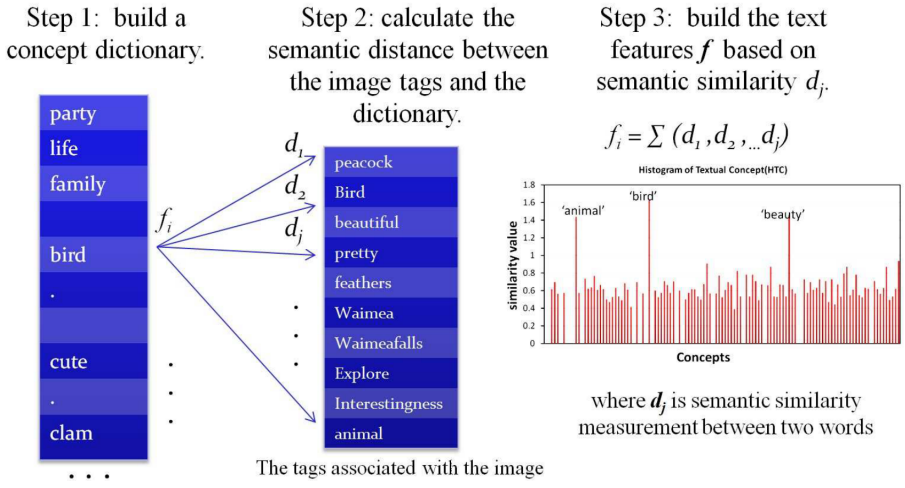


FIGURE 1.6 – Schéma de principe de l'algorithme HTC (appliqué au couple image/tags de la Figure 1.5).

Il est à noter que des tags tels que «peacock», «bird», «feathers», «animal» contribuent tous aux valeurs des classes de l'histogramme associées aux concepts

«animal» et «oiseau» selon une mesure de similarité sémantique, alors que des tags tels que «beautiful», «pretty», «interestingness» contribuent plutôt à la valeur de la classe associée au concept «cute». Ceci présente un clair avantage par rapport aux approches de type sac de mots pour lesquelles les liens entre concepts sont ignorés car ne reposant que sur les comptage de nombre d'occurences de mots. L'algorithme pour l'extraction de HTC est détaillé dans la Figure 1.7.

Algorithm 1: Histogram of Textual Concepts (HTC)

Input: Tag data $W = \{w_t\}$ with $t \in [1, \mathtt{T}]$, dictionary $D = \{d_i\}$ with $i \in [1, \mathtt{d}]$.
Output: Histogram f composed of values f_i with $0 \leq f_i \leq 1$, $i \in [1, \mathtt{d}]$.

- Preprocess the tags by using a stop-words filter.
- If the input image has no tags ($W = \emptyset$), return f with $\forall i \ f_i = 0.5$.[1]
- Do for each word $w_t \in W$:
  1. Calculate $dist(w_t, d_i)$, where $dist$ is a semantic similarity distance between w_t and d_i.
  2. Obtain the semantic matrix S as: $S(t, i) = dist(w_t, d_i)$.
- Calculate the feature f as: $f_i = \sum_{t=1}^{\mathtt{T}} S(t, i)$, and normalize it to [0 1] as: $f_i = f_i / \sum_{j=1}^{\mathtt{d}} f_j$.

FIGURE 1.7 – [1]Lorsque l'image ne comporte pas de tag, nous considérons que la valeur des classes de l'histogramme est de 0.5, et donc à mi-chemin entre une mesure de similarité sémantique de 0 (aucune relation avec le concept correspondant dans le dictionnaire), et 1 (similarité sémantique parfaite avec le concept correspondant dans le dictionnaire).

Les avantages de HTC sont multiples. Tout d'abord, pour un document textuel succinct tel que les tags d'une image, HTC permet une description des liens sémantique entre les tags des utilisateurs et les concepts textuels du dictionnaire. Ensuite, en cas de polysémie, HTC aide à désambiguïser les concepts textuels en fonction du contexte. Par exemple, le concept «bank» peut faire référence à un établissement financier mais également aux rives d'un cours d'eau. Cependant, si le mot «bank» est associé à une image représentant un établissement financier, des tags corrélés tels que «finance», «building», «money», etc., seront vraisemblablement utilisés, faisant ainsi une claire distinction entre le concept «bank» de la finance et celui de la rivière pour lesquels des tags corrélés seraient plutôt «water», «boat», «river», etc. De plus, en cas de synonymes, HTC renforcera le concept qui leur est associé puisque la mesure de similarité sémantique prend ceci en compte.

Le calcul de HTC nécessite un dictionnaire et une mesure de similarité sémantique appropriée aux concepts textuels. Ces choix doivent se faire en fonction du

contexte et des objectifs de l'application. Ainsi, nous avons proposé plusieurs variantes, notamment l'utilisation d'un dictionnaire contenant les concepts visuels à reconnaître pour la tâche de classification d'images associées à des tags. De plus, plusieurs mesures de similarité sémantiques ont été évaluées, reposant sur l'ontologie Wordnet représentant les relations lexicales et sémantiques entre concepts pour la langue anglaise [Mil95]. Enfin, afin d'associer aux relations sémantiques l'importance des mots dans les annotations textuelles, nous avons proposé une combinaison de HTC avec TF-IDF notée $cHTC$ telle que $cHTC[i] = HTC[i] + TFIDF[i]$ où i représente le $i^{ème}$ mot du dictionnaire et donc la $i^{ème}$ classe de l'histogramme HTC.

HTC et ses variantes ont été développées pour notre participation à la tâche "Photo annotation" d'ImageCLEF en 2011 et en 2012. L'objectif est de proposer des modèles pour annoter automatiquement un ensemble de 10 000 images selon 94 concepts visuels, en disposant d'un ensemble de 15 000 images pour l'entraînement. Les images, issues du site Flickr [1], sont accompagnées d'information textuelles dont les tags des utilisateurs. Chaque équipe participante peut proposer jusqu'à 5 soumissions pouvant ne reposer que sur des descripteurs textuels, que sur des descripteurs visuels, ou sur une approche multimodale. Cette tâche est décrite plus en détail dans la section 1.4.

En 2012, 18 équipes internationales ont participé avec un total de 80 soumissions dont 17 étaient exclusivement textuelles.

Notre meilleur modèle textuel basé sur HTC a obtenu la meilleure performance parmi toutes les soumissions textuelles, comme relaté dans la Figure 1.8. Le modèle de classification utilisant ces descripteurs HTC pour prédire les concepts visuels est SWLF, présenté dans la section 1.4.

Nos expérimentations ont montré que nos descripteurs HTC permettent d'améliorer nettement les performances par rapport à l'utilisation exclusive de descripteurs visuels, en particulier pour les concepts disposant de peu de données d'apprentissage. Ceci est illustré dans la Figure 1.9. Dans le premier exemple, la photo a été prise à l'intérieur d'un avion, ce qui n'est pas fréquent dans l'ensemble d'apprentissage, et donc les modalités visuelles n'ont pas permis de détecter le concept "airplane". Par contre, le texte associé à l'image contenant le tag «plane», notre descripteur HTC a réussi à capturer cette information et donc a augmenter le score du concept «airplane» qui lui est proche sémantiquement, ce qui a permis au classifieur multimodal de finalement faire la bonne prédiction de classe. Le même phénomène peut être observé dans le deuxième exemple avec le skateboard et le troisième exemple avec la pluie.

1. www.flickr.com

Team	RankMiAP		Team	RankGMiAP		Team	Rankmicro-F1	
LIRIS	1	0.3338	LIRIS	1	0.2771	LIRIS	1	0.4691
CEA LIST	3	0.3314	CEA LIST	2	0.2759	IMU	2	0.4685
IMU	4	0.2441	IMU	4	0.1917	CEA LIST	5	0.4452
CERTH	6	0.2311	CERTH	7	0.1669	MLKD	7	0.3951
MSATL	8	0.2209	MSATL	9	0.1653	CERTH	8	0.3946
IL	11	0.1724	IL	11	0.1140	IL	10	0.3532
BUAA AUDR	13	0.1423	BUAA AUDR	13	0.0818	URJCyUNED	11	0.3527
UNED	14	0.0758	UNED	14	0.0383	MSATL	13	0.2635
MLKD	15	0.0744	MLKD	15	0.0327	BUAA AUDR	14	0.2167
URJCyUNED	17	0.0622	URJCyUNED	17	0.0254	UNED	16	0.0864

FIGURE 1.8 – Résumé des résultats d'annotation pour les soumissions textuelles à la tâche "Photo Annotation" d'ImageCLEF 2012. Notre soumission (LIRIS) a obtenu la meilleure performance pour les trois mesures.

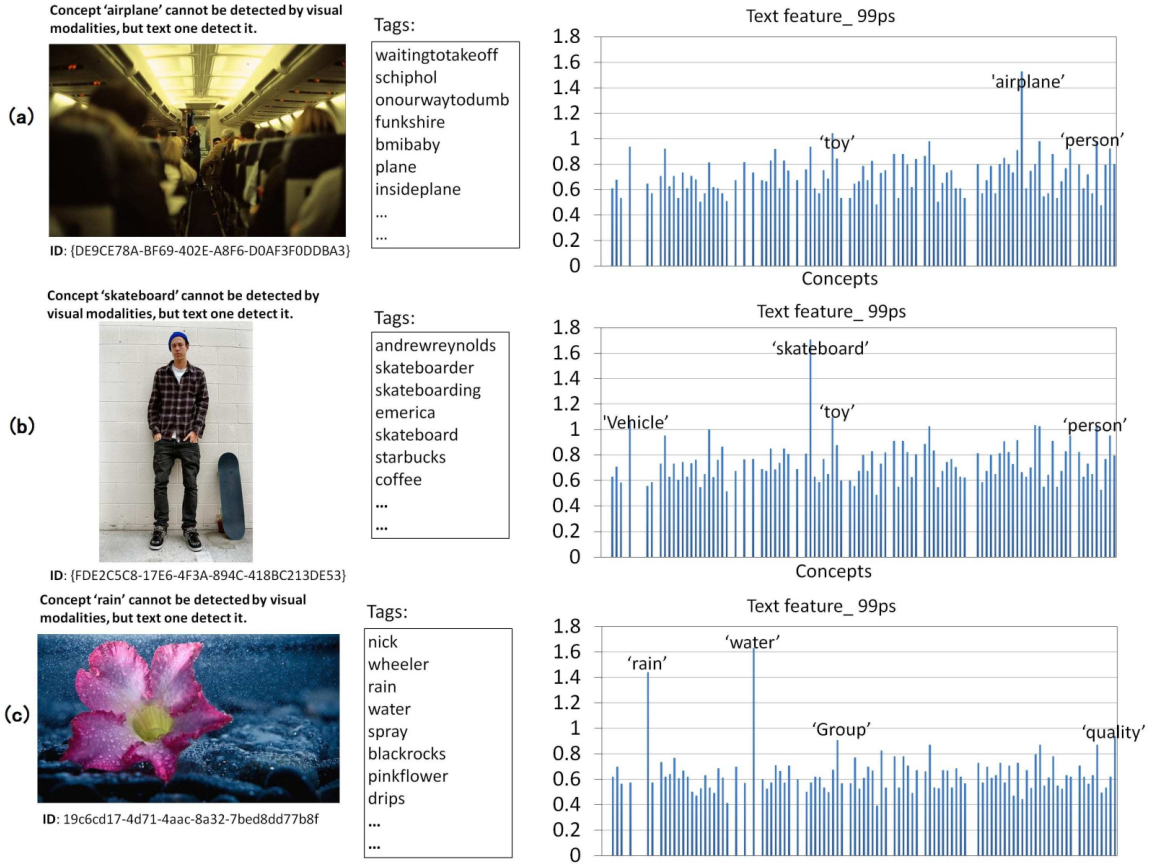


FIGURE 1.9 – La colonne de gauche présente les images de trois concepts visuels qui n'ont pas pu être identifiés par les modalités visuelles, mais dont les descripteurs HTC ont finalement pu mener à une prédiction correcte. La colonne centrale représente le texte associé aux images et la colonne de droite l'histogramme HTC utilisant le dictionnaire formé des 99 concepts visuels à reconnaître dans la tâche "Photo Annotation d'ImageCLEF 2012".

1.4 Fusion multimodale

Lorsque plusieurs sources d'information sont à disposition pour caractériser des données visuelles, il devient nécessaire de les combiner avec pour objectif d'en tirer le meilleur parti pour les concepts visuels à reconnaître [KAB+19]. Cette fusion peut être réalisée au niveau des descripteurs (fusion précoce) [PBAC+17], au niveau du score du classifieur (fusion tardive) [SWS05, TSBB+14], voire même à des niveaux intermédiaires, comme par exemple au niveau de noyaux [AQG07].

La fusion précoce a l'avantage d'être simple, consistant généralement à concaténer les descripteurs issus de différentes sources d'information afin d'obtenir un unique vecteur de représentation des données. Ses inconvénients sont que la dimension peut devenir très grande et donc mener à la malédiction de la dimension [BC61], et que combiner des descripteurs de différentes natures en une représentation très hétérogène peut être problématique et conduire à de faibles performances. De leur côté, les stratégies de fusion tardive, qui consistent à intégrer les scores délivrés par des classifieurs s'appuyant sur divers descripteurs, grâce à des règles de combinaison, telle que la somme, sont des alternatives souvent plus efficaces [WCCS04, TLN+03].

La méthode que nous avons proposée, SWLF pour "Selective Weighted Later Fusion", relève de cette dernière catégorie de fusion tardive, son objectif étant de combiner efficacement différentes sources d'information pour le problème de la classification d'images en particulier lorsqu'elles sont associées à une description textuelle.

Dans la mesure où une fusion tardive au niveau des scores des classifieurs est reconnue pour être une manière simple et efficace pour combiner des descripteurs de nature différente, SWLF s'appuie sur deux idées simples. Premièrement, le score de classification à partir d'un type de descripteur (classifieur expert) doit être pondéré en fonction de sa qualité intrinsèque pour le problème de classification en question. Deuxièmement, dans le cadre d'un scénario multi-labels où plusieurs concepts visuels peuvent être attribuées à une même image [HMQ16], différents concepts visuels peuvent nécessiter différents types de descripteurs pour permettre leur reconnaissance de manière efficace.

Par exemple, le concept «ciel» pourrait nécessiter des descripteurs globaux de couleur alors que le concept «rue» serait mieux caractérisé par des descripteurs basés sur des segments pour capturer les propriétés des lignes droites des immeubles. Le schéma de principe de l'algorithme SWLF est présenté dans la Figure 1.10.

Le principe de SWLF est le suivant. L'ensemble d'apprentissage initial est tout d'abord divisé en deux parties : un ensemble d'apprentissage et un ensemble de validation. Pour chaque concept visuel, un classifieur binaire (un contre tous), appelé

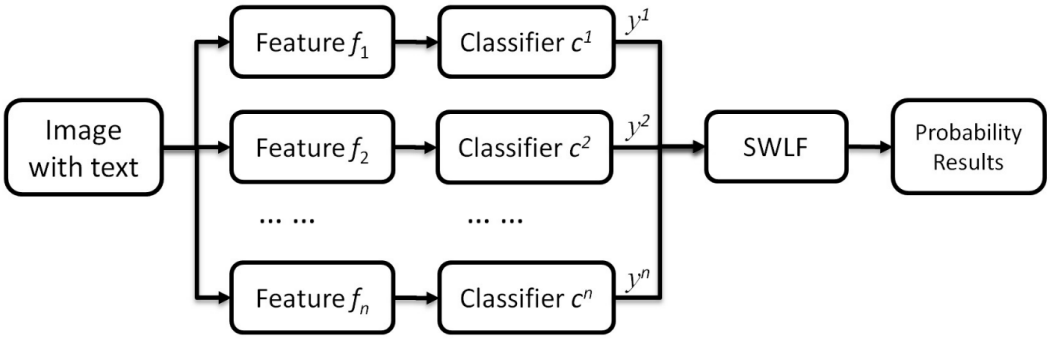


FIGURE 1.10 – Schéma de principe de l'algorithme SWLF. Pour chaque image et chaque concept, les tags associés aux images sont extraits et traités pour former les descripteurs textuels pour les classifieurs textuels. Parallèlement, les descripteurs visuels sont extraits pour alimenter les classifieurs visuels. Les classifieurs sont ensuite combinés pour prédire la présence d'un concept donné dans l'image d'entrée.

"expert" par la suite, est entraîné, pour chaque type de descripteur en utilisant l'ensemble d'apprentissage. Ainsi, pour chaque concept, le nombre d'experts générés est égal au nombre de types de descripteurs. La qualité d'un expert peut être évaluée au travers d'une métrique de performance calculée sur l'ensemble de validation. Dans nos travaux, la métrique utilisée est la précision moyenne interpolée (iAP). Dans ce cas, plus l'iAP est élevée pour un expert donné, et plus le poids donné au score de cet expert sera important dans la somme pondérée des scores pour la fusion tardive.

Concrètement, pour un concept visuel k, les métriques de qualité (iAP) produites par tous les experts sont d'abord normalisées et stockées dans les coefficients ω_k^i. Pour permettre une fusion tardive de tous les experts au niveau du score, la somme des scores pondérés est ensuite calculée selon l'équation(1.17).

$$score : z_k = \sum_{i=1}^{N} (\omega_k^i * y_k^i) \tag{1.17}$$

où y_k^i représente le score du $i^{ième}$ expert pour le concept k, et ω_k^i représente la performance iAP normalisée pour le descripteur f_i sur l'ensemble de validation.

En pratique, il s'avère que la fusion de tous les experts entraîne une baisse de performance qur l'ensemble des concepts visuels à reconnaître. Cela est dû au fait que certains descripteurs peuvent être bruités et/ou non pertinents pour certains concepts visuels, et donc perturber le processus d'apprentissage et diminuer la capacité de généralisation de l'expert sur de nouvelles données. Nous avons donc intégré à SWLF un processus de sélection des descripteurs inspiré de l'algorithme SFS (Sequential Forward Selection [PNK94]) dont le principe est de partir d'un ensemble

vide et d'ajouter à chaque étape le descripteur qui, ajouté à ceux sélectionnés précédemment, permet d'obtenir le meilleur score de classification. Dans le cas de SWLF, pour chaque concept visuel, tous les experts sont triés par ordre décroissant selon leur iAP. A une itération N donnée, seuls les N premiers experts sont utilisés pour la fusion et leurs performances sont alors évaluées sur l'ensemble de validation. N continue d'augmenter jusqu'à ce que le taux de classification global mesuré en terme de MiAP (Mean iAP) commence à décroître. L'algorithme complet de SWLF est détaillé dans la Figure 1.11.

Algorithm 2: Selective Weighted Late Fusion (SWLF)

Input: Training dataset T (of size N_T) and validation dataset V (of size N_V).
Output: Set of N experts for the K concepts $\{C_k^n\}$ and the corresponding set of weights $\{\omega_k^n\}$ with $n \in [1, N]$ and $k \in [1, K]$.
Initialization: $N = 1$, $MiAP_{max} = 0$.

- Extract M types of features from T and V
- For each concept $k = 1$ to K
  - For each type of feature $i = 1$ to M
    1. Train the expert C_k^i using T
    2. Compute ω_k^i as the iAP of C_k^i using V
  - Sort the ω_k^i in descending order and denote the order as j^1, j^2, ..., j^M to form $W_k = \{\omega_k^{j^1}, \omega_k^{j^2}, ..., \omega_k^{j^M}\}$ and the corresponding set of experts $E_k = \{C_k^{j^1}, C_k^{j^2}, ..., C_k^{j^M}\}$
- For the number of experts $n = 2$ to M
  - For each concept $k = 1$ to K
    1. Select the first n experts from E_k : $E_k^n = \{C_k^1, C_k^2, ..., C_k^n\}$
    2. Select the first n weights from W_k : $W_k^n = \{\omega_k^1, \omega_k^2, ..., \omega_k^n\}$
    3. For $j = 1$ to n : Normalise $\omega_k^{j'} = \omega_k^j / \sum_{i=1}^n \omega_k^i$
    4. Combine the first n experts into a fused expert, using the *weighted score* rule through Equation (9): $z_k = \sum_{j=1}^n \omega_k^{j'}.y_k^j$ where y_k^j is the output of C_k^j
    5. Compute $MiAP_k^n$ of the fused expert on the validation set V
  - Compute $MiAP = 1/K. \sum_{k=1}^K MiAP_k^n$
  - If $MiAP > MiAP_{max}$
    * Then $MiAP_{max} = MiAP$, $N = n$
    * Else break

FIGURE 1.11 – Algorithme de notre méthode de fusion multimodale tardive «Selective Weighted Late Fusion» (SWLF).

Finalement, le schéma de principe de notre approche de classification multimodale d'images avec texte compagnon est donné dans la Figure 1.12.

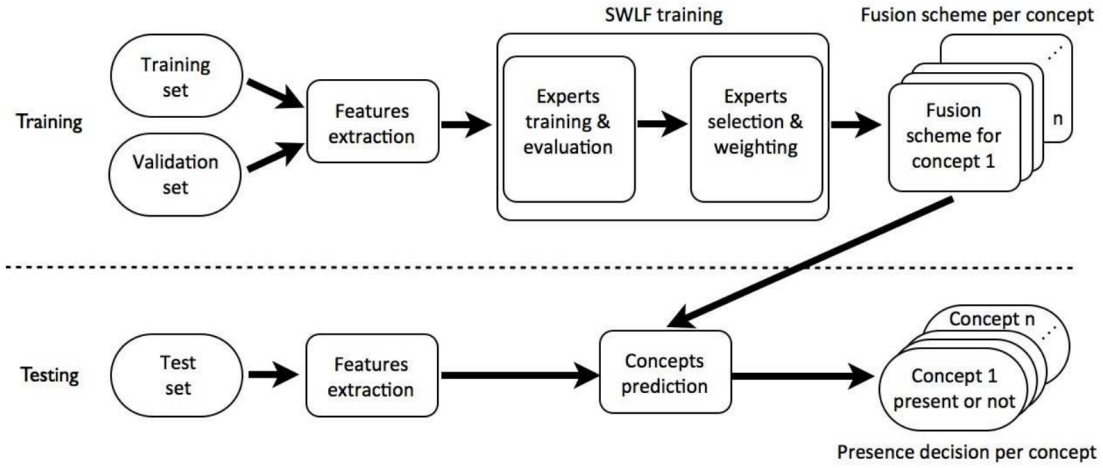


FIGURE 1.12 – Schéma de principe de notre approche multimodale pour la classification d'images.

Nos expérimentations ont été conduites sur la collection d'images MIR FLICKR [HL08, HTL10] qui a été utilisée pour les tâches "Photo Annotation" à ImageCLEF 2011 et 2012 auxquelles nous avons participé.

Notre approche multimodale s'est appuyée sur 11 descripteurs textuels issus de notre algorithme HTC présenté à la section 1.3 et 24 descripteurs visuels locaux et globaux décrivant les propriétés de couleur, texture et forme des images.

Les classieurs experts utilisés par SWLF sont des Support Vector Machines (SVM) [Vap95] qui offrent l'avantage d'être généralement efficaces en terme de complexité calculatoire et de performance de classification. Un expert SVM a été entraîné pour chaque concept et chaque descripteur, comme décrit dans l'algorithme de la Figure 1.11. Comme suggéré par Zhang et al. [ZMLS07], des noyaux de type χ^2 ont été utilisés pour les descripteurs visuels sous forme d'histogramme, et de type Radial Basis Function (RBF) pour les autres descripteurs. Les hyper-paramètres ont été optimisés sur l'ensemble de validation.

Pour la tâche "Photo Annotation" à ImageCLEF2012, les participants avaient pour objectif d'élaborer des méthodes afin d'annoter automatiquement un ensemble de test de 10 000 images selon 94 concepts visuels. Un ensemble d'apprentissage de 8 15 000 images était fourni. La particularité de cette tâche est que les images sont associées à un texte compagnon contenant notamment les tags des utilisateurs issus du site Flickr [2].

Chaque équipe participante a la possibilité de proposer jusqu'à 5 soumissions. Celles-ci peuvent reposer uniquement sur la modalité textuelle, uniquement sur la

2. www.flickr.com

modalité visuelle ou être multimodale (textuelle+visuelle).

Les performances ont été évaluées selon les mesures Mean Interpolated Average Precision (MiAP), Geometric Mean interplated Average Precision (GMiAP) et F1.

18 équipes internationales ont participé et soumis un total de 80 soumissions parmi lesquelles 17 étaient purement textuelles, 28 purement visuelles et 35 multi-modales. Nous avons pour notre part soumis 5 soumissions : 2 textuelles, 1 visuelle et 2 multi-modales.

Les résultats pour les 18 équipes sont donnés dans la Figure 1.13. Ils indiquent que l'une de nos soumissions multimodales (LIRIS) a obtenu la première place parmi les 80 soumissions pour les trois mesures de performance. Cela montre que notre modèle SWLF, qui sélectionne et combine le meilleur ensemble de classifieurs experts tout en optimisant la performance globale, choisie ici comme la MiAP, est doté d'une très bonne capacité de généralisation sur les nouvelles données et est particulièrement efficace pour le problème d'annotation automatique d'images dans un scénario multi-labels, nécessitant la fusion de descripteurs textuels et visuels. L'ensemble des expérimentations plus détaillées peuvent être trouvées dans [LDC+13] et [LDTC14].

Team	Rank	MiAP	Feature	Team	Rank	GMiAP	Feature	Team	Rank	micro-F1	Feature
LIRIS	1	0.4367	M	LIRIS	1	0.3877	M	LIRIS	1	0.5766	M
DMS-SZTAKI	3	0.4258	M	DMS-SZTAKI	3	0.3676	M	DMS-SZTAKI	3	0.5731	M
CEA LIST	6	0.4159	M	CEA LIST	5	0.3615	M	NII	6	0.5600	V
ISI	7	0.4136	M	ISI	7	0.3580	M	ISI	7	0.5597	M
NPDPILIP6	16	0.3437	V	NPDPILIP6	16	0.2815	V	MLKD	16	0.5534	V
NII	22	0.3318	V	NII	21	0.2703	V	CEA LIST	20	0.5404	M
CERTH	28	0.3210	M	MLKD	28	0.2567	V	CERTH	26	0.4950	M
MLKD	29	0.3185	V	CERTH	29	0.2547	M	IMU	30	0.4685	T
IMU	36	0.2441	T	IMU	35	0.1917	T	KIDS NUTN	34	0.4406	M
UAIC	38	0.2359	V	UAIC	39	0.1685	V	UAIC	35	0.4359	V
MSATL	41	0.2209	T	MSATL	42	0.1653	T	NPDPILIP6	37	0.4228	V
IL	46	0.1724	T	IL	45	0.1140	T	IL	49	0.3532	T
KIDS NUTN	47	0.1717	M	KIDS NUTN	49	0.0984	M	URJCyUNED	50	0.3527	T
BUAA AUDR	52	0.1423	V	BUAA AUDR	51	0.0818	V	PRA	54	0.3331	V
UNED	55	0.1020	V	UNED	55	0.0512	V	MSATL	57	0.2635	T
DBRIS	58	0.0976	V	DBRIS	57	0.0476	V	BUAA AUDR	58	0.2592	M
PRA	65	0.0900	V	PRA	66	0.0437	V	UNED	66	0.1360	V
URJCyUNED	77	0.0622	V	URJCyUNED	77	0.0254	V	DBRIS	69	0.1070	V

FIGURE 1.13 – Résumé des résultats d'annotation pour la meilleure soumission des 18 équipes pour la tâche "Photo Annotation" à ImageCLEF 2012. Pour la colonne "Feature", T indique qu'il s'agit d'une soumission Textuelle, V Visuelle et M Multimodale.

1.5 Conclusion

Ce chapitre retrace nos principales contributions dans le domaine de la classification d'images.

Nous avons en particulier proposé une nouvelle approche de représentation parcimonieuse des images adaptée à la classification grâce à l'ajout d'un terme discriminatif dans la fonction objectif de représentation parcimonieuse afin de permettre l'apprentissage d'un dictionnaire à la fois reconstructif et discriminatif. Ce terme discriminatif flexible peut par exemple être la mesure discriminative de Fisher ou encore le score d'un classifieur (SVM dans nos expérimentations). Les expérimentations que nous avons menées sur le jeu de données SIMPLIcity ont montré que notre approche reconstructive a permis d'améliorer de manière significative les performances de classification par rapport à celles d'un classifieur standard de type SVM utilisant directement en entrée les descripteurs d'images. De plus, notre approche reconstructive et discriminative est plus efficace qu'une approche purement reconstructive, ce qui montre qu'ajouter un terme discriminatif pour l'élaboration de la représentation parcimonieuse est plus adapté à la classification d'images.

D'autre part, nous avons également proposé un nouveau descripteur textuel, «Histogram of Textual Concepts» (HTC) afin d'exploiter l'information textuelle associée aux images, tels que des légendes ou des tags d'utilisateurs. Elle repose sur la similarité sémantique entre les mots associés à l'image et un dictionnaire de concepts. Afin de combiner efficacement les différentes modalités textuelles et visuelles, un nouveau schéma de fusion tardive sélective a été introduit, «Selective Weighted Late Fusion» (SWLF) qui sélectionne itérativement les meilleurs descripteurs et pondère le score des classifieurs experts associés pour chaque concept à identifier. Nos expérimentations ont été réalisées sur le jeu de données MIR FLICKR utilisé dans le cadre de la tâche «Photo Annotation» à ImageCLEF. Une de nos soumissions à ce challenge en 2012 a obtenu la meilleure performance parmi les 80 soumissions de 18 équipes. L'ensemble de nos expérimentations ont montré que notre descripteur textuel HTC permet d'augmenter de manière significative les performances de classifieurs d'images en particulier pour les catégories disposant de peu de données d'apprentissage, et que la fusion de classifieurs experts par notre méthode SWLF sélectionnant et combinant un ensemble de classifieurs experts tout en optimisant une métrique de performance (la précision moyenne par exemple), montre une très bonne capacité de généralisation sur de nouvelles données en fusionnant de manière efficace les informations visuelles et textuelles.

Ces travaux ont été réalisés à une période charnière : à partir de 2012, l'avènement de l'apprentissage profond (ou «deep learning») a provoqué une révolution dans le domaine. Avec en particulier la capacité d'apprendre au sein du même modèle une représentation de l'image grâce aux couches de convolution et le moyen de discriminer les classes grâce aux couches entièrement connectées, ces modèles convo-

lutifs ont permis d'obtenir un gain de performances extrêmement important par rapport aux modèles utilisés précédemment reposant sur le paradigme «extraction de descripteurs puis apprentissage d'un classifieur». Ces modèles profonds continuent d'être la référence en classification d'images, mais ont ouvert la voie à de nouvelles problématiques, en particulier du fait de la nécessité de disposer d'un nombre très important de données pour réaliser l'apprentissage. Une solution possible à ce problème est apportée dans le chapitre suivant dans le cas de la détection d'objets dans les images.

Chapitre 2

Détection d'objets dans les images

2.1 Introduction

Au delà de la classification d'images permettant d'identifier les concepts visuels dans les images, il peut être nécessaire dans certaines circonstances de localiser les objets dans cette image. La difficulté est alors d'avoir à notre disposition un nombre suffisant d'images annotées manuellement avec des boîtes englobantes précisant la nature et la localisation des objets afin de réaliser l'apprentissage d'un modèle de détection. En effet, il est beaucoup plus laborieux et beaucoup moins fiable d'annoter des boîtes englobantes plutôt que d'attribuer une étiquette globale à l'image. C'est la raison pour laquelle les jeux de données contenant des images avec des annotations au niveau de boîtes englobantes sont beaucoup moins volumineux que ceux contenant des images annotées globalement.

Dans ce contexte, nous nous sommes intéressés au problème de la détection d'objets faiblement supervisée et semi-supervisée. Dans le premier cas, le but est alors de reconnaître et de localiser des objets dans les images, n'ayant à notre disposition durant la phase d'apprentissage que des annotations au niveau global des images (pas de boîte englobante), alors que dans le deuxième cas, seule une partie des catégories à reconnaître possède des annotations au niveau des objets. Nous avons donc proposé deux modèles de détection d'objets adaptés à ces deux situations.

Ces travaux ont été réalisés avec le doctorant Yuxing Tang, notamment dans le cadre du projet CHIST-ERA Visen, et ont été publiés dans [R1, R4, C4, C7] (numérotation correspondant aux publications listées dans le rapport d'activité en fin de document).

2.2 Apprentissage faiblement supervisé de modèles à parties déformables

L'approche «Deformable Part-based Models» (DPM) [FGMR10] et ses variantes [GFM11, AL12, RR13] ont été parmi les plus efficaces dans le domaine de la détection d'objets supervisée, pendant une longue période, en particulier sur le difficile jeu de données PASCAL VOC [EGW+10].

DPM représente un objet avec un filtre racine holistique couvrant approximativement l'ensemble de l'objet, et plusieurs filtres haute résolution capturant de plus petites propriétés d'apparence locale correspondant aux parties de l'objet. Les déformations sont également caractérisées par des liens connectant les différentes parties. Dans le cadre du modèle standard supervisé, le filtre racine est initialisé avec la boîte englobante contenant l'objet, et est autorisé à se déplacer dans un voisinage très proche pour maximiser le score du filtre. Les localisations des parties de l'objet sont toujours traitées comme une information latente du fait de l'indisponibilité dans la grande majorité des cas des annotations concernant ces parties. Un SVM latent (LSVM) est en général utilisé pour apprendre les déformations de l'objet.

Dans [PL11], le modèle DPM supervisé a été transformé en DPM faiblement supervisé sans annotation au niveau des objets, en traitant la position du filtre racine et des filtres de parties de manière entièrement latente et en apprenant les detecteurs d'objets structurés à partir de l'image entière. La position du filtre racine est initialisée aléatoirement à partir d'une fenêtre ayant au minimum 40% de recouvrement avec l'image d'apprentissage, et son rapport de forme est initialisé approximativement à la moyenne des rapports de forme des exemples d'apprentissage. Cependant, la taille, la position et le rapport de forme du filtre racine initial ont une grande influence sur les performances de localisation finale [DT05, FGMR10, PL11]. En effet, avec une initialisation aléatoire, le détecteur d'objets est susceptible d'apprendre des modèles erronés d'autre classes ou de régions de l'arrière-plan, menant à une efficacité réduite de détection sur les données de test.

Estimation initiale des objets

Ainsi, afin de réduire l'écart de performances entre le DPM faiblement et celui totalement supervisé, nous avons proposé une amélioration de l'approche DPM faiblement supervisée, en insistant sur l'importance de la position et de la taille du filtre racine initial spécifique à la classe. Tout d'abord, un ensemble de candidats est calculé, ceux-ci représentant les positions possibles de l'objet pour le filtre racine

initial, en se basant sur une mesure générique d'objectness (par region proposals) pour combiner les régions les plus saillantes et potentiellement de bonne qualité. Ensuite, nous avons proposé l'apprentissage du label des classes latentes de chaque candidat comme un problème de classification binaire, en entraînant des classifieurs spécifiques pour chaque catégorie afin de prédire si les candidat sont potentiellement des objets cible ou non. De plus, nous avons amélioré la détection en incorporant l'information contextuelle à partir des scores de classification de l'image. Enfin, nous avons élaboré une procédure de post-traitement permettant d'élargir et de contracter les régions fournies par le DPM afin de les adapter efficacement à la taille de l'objet, augmentant ainsi la précision finale de la détection.

La première étape consiste donc à identifier les régions initiales pouvant potentiellement contenir des objets cibles dans les images d'entraînement positives avec une annotation uniquement au niveau de l'image globale. La démarche est présentée dans la Figure 2.1.

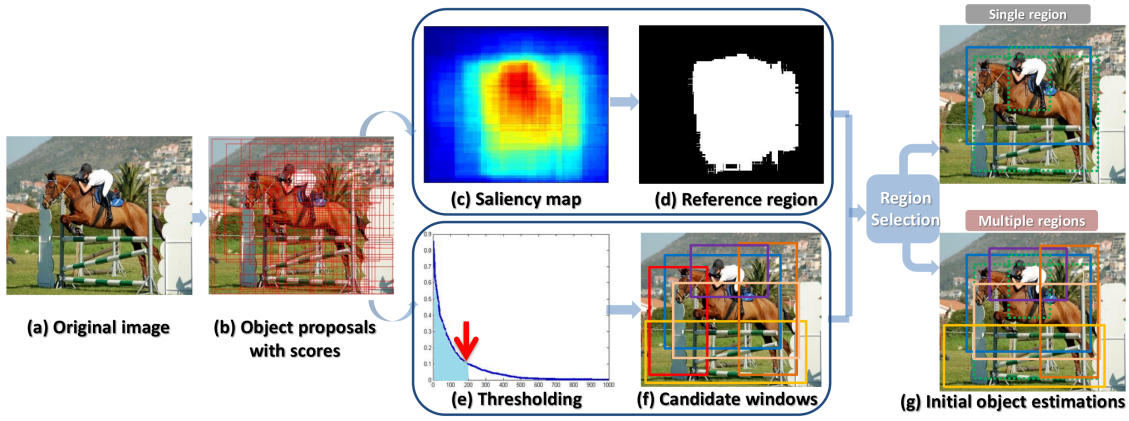


FIGURE 2.1 – Schéma de principe pour l'estimation initiale des objets. Pour une image d'entrée (a), les propositions d'objets (rectangles englobants) (b) sont calculées ainsi que leur score correspondant à la probabilité de contenir une objet. (c) représente la carte de saillance estimée à partir de (b), et (d) est la région de référence obtenue par seuillage de (c). (f) représente l'ensemble de régions candidates fournies par Selective Search et sélectionnées par la méthode non-maximum suppression (NMS). Dans l'image supérieure de (g), correspondant à la stratégie de sélection d'une région unique, la fenêtre bleue représente l'estimation d'objet initiale obtenue en fusionnant les informations fournies par (d) et (f). L'image inférieure de (g) correspond à la stratégie de sélection de régions multiples. Les fenêtres de couleur avec des traits pleins sont les estimations des régions susceptibles de contenir les objets de l'image. Pour chaque image de (g), les fenêtres à traits verts en pointillés sont les boîtes englobantes de la vérité terrain pour les catégories «person» et «horse».

A partir d'une image d'entrée I (Figure 2.1(a)), la méthode Selective Search [USGS13] est tout d'abord appliquée afin de sélectionner les meilleures n régions

$W = \{w_1, w_2, \ldots, w_n\}$ ainsi que leurs scores associés $S = \{s_1, s_2, \ldots, s_n\}$ indiquant la probabilité de la région de recouvrir un objet (Figure 2.1(b)) Cette méthode permet généralement de capturer tous les objets possibles dans l'image. Cependant, les régions avec les scores les plus élevés ne sont pas toujours les meilleurs choix [SSX12] car elles peuvent être bruitées par des parties importantes de l'arrière-plan, ou ne recouvrir un objet que partiellement. Ainsi, afin de sélectionner un ensemble fiable de régions à partir de W, nous avons proposé une procédure de sélection récursive (Figure 2.1(c)-(g)). Afin d'éviter une initialisation du filtre racine dans l'arrière-plan de l'image, ce qui compromettrait la détection de l'objet, nous avons proposé de prendre en compte l'information de saillance qui aura plutôt tendance à mettre en évidence les objets en avant-plan [GDBBP16, CBPZBA17]. La région de référence R (Figure 2.1(d)) est alors obtenue en seuillant et fusionnant la carte de saillance M (Figure 2.1(c)). La valeur de la carte de saillance M pour un pixel $I(i, j)$ est obtenue en additionnant les scores des régions recouvrant ce pixel :

$$M(i, j) = \sum_{k=1}^{n} M_k(i, j) \tag{2.1}$$

où

$$M_k(i, j) = \begin{cases} s_k & \text{if } I(i, j) \in w_k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Le score délivré par la méthode Selective Search (score d'objectness) correspond grossièrement à la probabilité d'avoir un objet à l'intérieur de la région correspondante. Afin de prendre en compte cette information intéressante, nous sélectionnons 200 régions parmi les n régions ayant les scores les plus élevés (Figure 2.1(e)). Pour éviter d'obtenir des régions candidates très similaires, la méthode non-maximum suppression (NMS) est appliquée, permettant d'obtenir un ensemble plus pertinents de l régions candidates $\hat{W} = \{\hat{w_1}, \hat{w_2}, \ldots, \hat{w_l}\}$ ainsi que leurs scores associés $\hat{S} = \{\hat{s_1}, \hat{s_2}, \ldots, \hat{s_l}\}$ (Figure 2.1(f)).

Disposant de la région de référence R indiquant la ou les régions les plus saillantes de l'image, ainsi que des régions candidates $\hat{W}$ et de leurs scores $\hat{S}$, leur fusion peut permettre d'obtenir une information intéressante pour identifier la position des objets cibles. Nous avons proposé deux approches pour cela, selon que les images soient susceptibles de contenir un ou plusieurs objets.

Une première stratégie consiste à identifier une région unique w^* à partir de $\hat{W}$ susceptible de contenir l'objet cible. Intuitivement, nous nous attendons à ce que cette estimation de région couvre autant que possible la région saillante de

référence R et possède également un score d'objectness relativement élevé. Ainsi, l'estimation de la boîte englobante initiale de l'objet et son score d'objectness (w^*, s^*) (Figure 2.1(g), image supérieure) peuvent être obtenus en optimisant la fonction suivante :

$$(w^*, s^*) = \underset{\hat{w}_i \in \hat{W}, \hat{s}_i \in \hat{S}}{\operatorname{argmax}} \left[\alpha \hat{s}_i + (1-\alpha) \frac{area(R \cap \hat{w}_i}{area(R \cup \hat{w}_i} \right], i \in [1, l] \tag{2.3}$$

où α permet de contrôler l'influence du score d'objectness.

Cette première stratégie d'initialisation tend à sélectionner une région relativement grande pouvant contenir la partie la plus saillante de l'image. Cela peut produire de bons détecteurs d'objets DPM dans un cas faiblement supervisé, lorsque l'image ne contient que très peu d'objets. Par exemple, cette stratégie dans le cas de l'image supérieure de la Figure 2.1(g) permet d'obtenir la fenêtre bleue qui sera utilisée comme initialisation du filtre racine à la fois pour la catégorie «horse» et la catégorie «person».

Dans le cas où les images peuvent contenir plusieurs objets dispersés dans l'image (par exemple 2,5 objets en moyenne pour le jeu de données PASCAL VOC 2007), les détecteurs DPM peuvent être améliorés en fournissant plusieurs estimations d'objets pour l'initialisation des filtres racine, plutôt que d'entraîner les détecteurs d'objets avec une unique région par image.

Ainsi, pour chaque image, l'objectif de la seconde stratégie, l'initialisation de régions multiples, est alors de sélectionner un petit nombre d'estimations d'objets pouvant également capturer des objets plus petits et dispersés. Un critère semblable à celui exprimé dans l'équation (2.3) avec la fonction de score, peut alors être utilisé. Cependant, au lieu de sélectionner uniquement la fenêtre avec le score maximal, les Q meilleures fenêtres W^* sont alors retenues pour chaque image.

Après avoir généré plusieurs estimations d'objets pour chaque image, l'étape suivante est d'identifier le label de classe de chaque estimation, en ne disposant que des labels fournis uniquement globalement à l'image. Par exemple, dans l'image inférieur de la Figure 2.1(g), les fenêtres de couleur avec lignes pleines sont associées aux labels «horse» et «person». Cependant, pour le moment, nous ignorons lequel de ces objets est à l'intérieur de ces boîtes englobantes.

Apprentissage des classes d'objets latentes par classification de régions

Pour chaque image positive de l'ensemble d'apprentissage, nous avons donc généré Q estimations d'objets avec la stratégie d'initialisation de régions multiples. Considérons une catégorie d'objets, «horse» par exemple, qui dispose de P images d'apprentissage positives. Cela correspond donc à un total de $z = P * Q$ estimations d'objets. Evidemment, certaines de ces estimations correspondent vraisemblablement à d'autres catégories («person», «sheep», ... par exemple), à des parties d'objets ou encore à l'arrière-plan de l'image. Nous considérons que ces labels de classes sont une information latente et proposons de reformuler ce problème d'apprentissage de classes latentes en un problème de classification consistant à classer grossièrement les estimations d'objets en catégorie d'objet cible ou catégorie non-cible (autre classe, parties d'objet ou arrière-plan).

Nous avons utilisé les descripteurs de réseaux de neurones convolutifs (CNN) pour représenter les régions (estimations d'objets), de manière similaire à R-CNN [GDDM14] (sortie de la couche fc6 (vecteur de dimension 4096) d'un réseau AlexNet [KSH12] pré-entrainé sur ImageNet [RDS$^+$15]). Cependant, dans notre cas, l'ajustement (ou finetuning) sur le jeu de données cible du réseau CNN pré-entraîné n'est pas possible dans la mesure où l'annotation au niveau des objets n'est pas disponible pour les jeux de données faiblement annotés (annotation uniquement au niveau de l'image globale).

Considérons l'apprentissage d'un détecteur pour la catégorie d'objets «horse». Pour toutes les P images d'entraînement positives pour la catégorie «horse», z estimations d'objets sont générées. Or, seule une partie de ces z régions contiennent effectivement un objet cible «horse», les autres contenant d'autres types d'objets, des parties d'objets, voire l'arrière-plan. Les catégories latentes de ces régions sont apprises par une classification de régions.

Tout d'abord, un classifieur SVM linéaire [CL11] «horse» est entraîné en utilisant les images labellisées «horse» comme exemples d'apprentissage positifs, et celles «non horse» comme exemples négatifs. Un vecteur de descripteurs CNN de taille 4096, comme mentionné précédemment, est calculé sur les images entières. Ensuite, le classifieur «horse» entraîné est appliqué sur les z estimations d'objets dans les images d'entraînement positives. En seuillant les scores fournis par le SVM, un sous-ensemble z' des régions est obtenu à partir des z estimations initiales. Ces z' régions sont alors considérées comme réprésentant la catégorie cible «horse», pouvant donc être traitées comme exemples d'entraînement positifs pour un détecteur

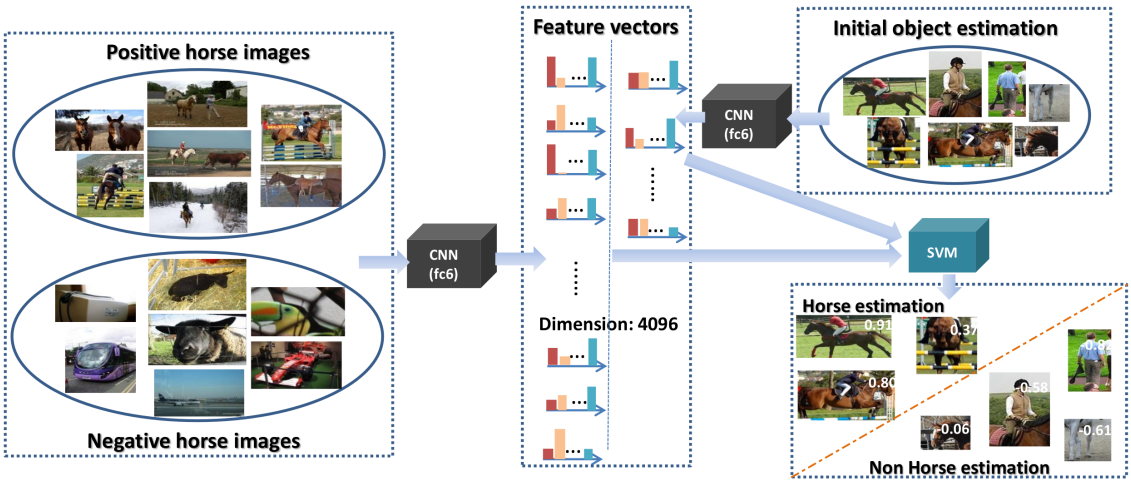


FIGURE 2.2 – Illustration de l'approche d'apprentissage de classes latentes, dans le cas de la catégorie «horse». Pour chaque catégorie d'objets, un classifieur de type SVM linéaire est entraîné avec des descripteurs CNN (sortie de la couche fc6 du CNN). Les estimations d'objets des images d'apprentissage positives de cette catégorie sont affectées du score fourni par le SVM. Les régions avec les scores les plus élevés sont alors sélectionnées pour représenter les objets de cette catégorie («horse» contre «non horse» dans cet exemple).

«horse». Cette procédure est illustrée dans la Figure 2.2. Si K catégories doivent être détectées, alors cette procédure est répliquée de la même manière pour les K classifieurs associées à ces catégories.

Apprentissage du DPM faiblement supervisé

L'apprentissage du modèle déformable par parties que nous avons proposé pour la détection d'objets faiblement supervisée «M-WDPM» (Multiple regions initialized Weakly supervised DPM) est réalisé de la façon suivante. De manière similaire à [FGMR10], chaque filtre racine dans une image d'entraînement positive est initialisé avec la boîte englobante obtenue avec la stratégie d'initialisation de régions multiples. La taille et le rapport de forme du filtre racine du DPM sont déterminées à partir de la moyenne des tailles et rapports de formes des estimations d'objets. Le filtre racine est autorisé à se déplacer dans un petit voisinage afin de maximiser le score du filtre de manière à compenser l'imprécision éventuelle des estimations des boîtes englobantes. La représentation d'une image est obtenue par la méthode «DeepPyramid» [GIDM15]. La carte de descripteurs correspond à la sortie de la $5^{ème}$ couche de convolution (conv5) qui possède 256 canaux de descripteurs. Chaque image (ou région) est représentée avec une pyramide de descripteurs à 7 niveaux. Pour l'entraînement, les estimations d'objets selectionnés sont traitées comme des

exemples d'apprentissage positifs et des fenêtres sélectionnées aléatoirements dans des images négatives constituent nos exemples d'apprentissage négatifs. Nous avons choisi un modèle DPM formé de 3 composants et 8 parties, comme suggéré dans [GIDM15]. Pour la phase de test, une approche s'appuyant sur une fenêtre glissantes dans l'image est utilisée.

Afin d'améliorer la prédiction, nous avons proposé une étape d'ajustement du score des boîtes englobantes à l'aide d'un classifieur CNN. La fonction d'ajustement, pour une fenêtre donnée, est la suivante :

$$s^i_{det} = \kappa s^i_{M-WDPM} + (1-\kappa) s^i_{cls}, \quad i \in [1, K] \tag{2.4}$$

où $0 \leq s^i_{M-WDPM} \leq 1$ est le score de détection normalisé du DPM du $i^{ème}$ détecteur et $0 \leq s^i_{cls} \leq 1$ est le score de classification softmax de la $i^{ème}$ catégorie. κ est un hyper-paramètre utilisé pour ajuster l'influence des deux scores. La prédiction finale est obtenue en seuillant la valeur de s^i_{det}.

L'apprentissage du classifieur CNN mentionné précédemment est réalisé par ajustement (finetuning) sur notre jeu de données d'un CNN pré-entraîné avec des annotations au niveau des images. La dernière couche softmax à 1000 neurones a été remplacée par une nouvelle couche softmax à K neurones. Tous les autres paramètres ont été conservés.

Un exemple de cette procédure de réajustement du score de détection des boîtes englobantes est donné Figure 2.3.

Enfin, la dernière étape que nous avons proposé consiste à post-traiter les boîtes engloblantes détectées. En effet, très souvent, les boîtes englobantes générées par les détecteurs DPM sont trop grandes (resp. petites) lors de la détection de très petits (resp. très grands) objets en raison des limitations de taille imposées par la taille du filtre racine et de l'échelle de la pyramide des descripteurs. Afin d'améliorer la localisation et d'obtenir une prédiction plus précise de la boîte englobante, une procédure d'élargissement et de contraction est appliquée à la boîte de manière à recouvrir au maximum l'objet. Cela est réalisé par une amélioration de la méthode proposée dans [KTJ06], et le principe général est d'augmenter la largeur et la hauteur de la boîte englobante d'origine de 120% et de calculer l'énergie des contours (à partir des gradients) dans cette boîte augmentée. Ensuite, à partir de la position du centroïde d'énergie, la nouvelle boîte est élargie dans les 4 directions jusqu'à ce que qu'un total de 98% de l'énergie de contour soit atteint. Ce post-traitement permet non seulement de supprimer les parties d'arrière-plan (lorsqu'ils sont relativement uniformes), mais également d'augmenter la taille de la boîte pour mieux couvrir

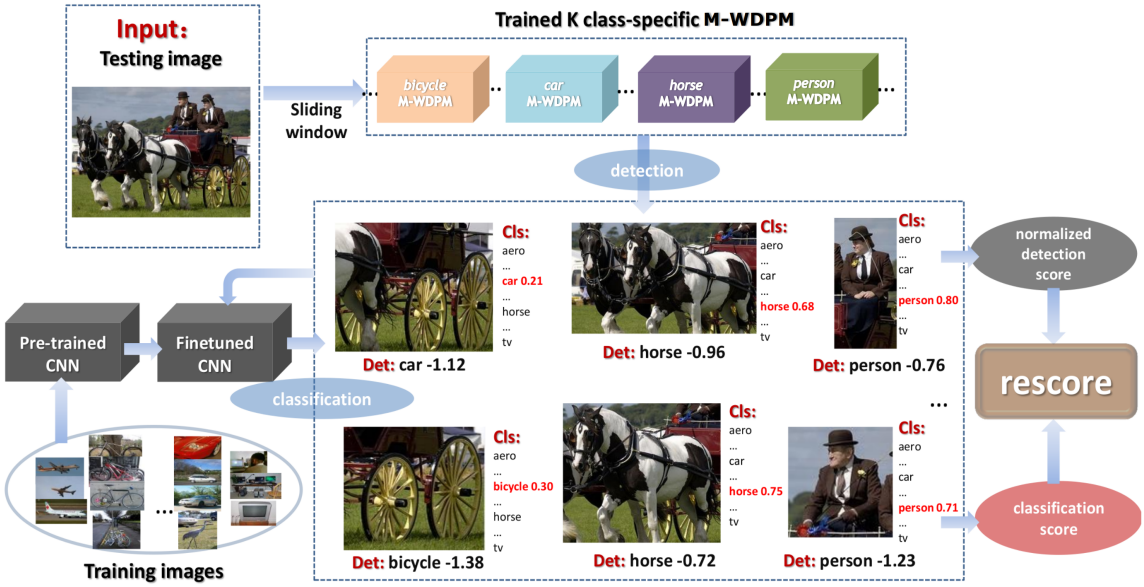


FIGURE 2.3 – Illustration de la procédure d'ajustement du score de détection des boîtes englobantes en utilisant M-WDPM et un classifieur CNN softmax. Pour une image test, K (nombre de classes dans le jeu de données cible) M-WDPMs spécifiques aux classes sont appliqués en utilisant une fenêtre glissante. Pour chaque fenêtre détectée par M-WDPM, le score de détection normalisé est combiné avec le score du classifieur softmax de la catégorie détectée. Dans cet exemple, les objets «car» et «bicycle» détectés avec erreur sont finalement rejetés après l'étape d'ajustement du score.

les objets détectés. Quelques exemples de ce post-traitement sont donnés dans la Figure 2.4.

Evaluation expérimentale

Notre modèle M-WDPM a été évalué sur le jeu de données PASCAL VOC 2007 [EEVG+15] qui contient 9963 images avec 20 catégories d'objets, divisé en un ensemble d'apprentissage (2501 images), de validation (2510 images) et de test (4952 images). Ce jeu de données est difficile en raison d'importantes similarités inter-classes, de grandes variances intra-classes, d'arrière-plans chargés et de variations importantes d'échelles. Pour cette tâche, nous avons uniquement utilisé les annotations des catégories au niveau de l'image globale.

Les résultats de détection sont donnés dans la Figure 2.5. Nous avons évalué trois variantes de notres approche : «M-WDPM-HOG» n'utilisant pas d'ajustement du score de détection et reposant sur une pyramide de descripteurs HOG [DT05], «M-WDPM-Deep» n'utilisant pas d'ajustement du score de détection et reposant sur une pyramide de descripteurs CNN, et «M-WDPM-rescore», notre méthode com-

FIGURE 2.4 – Exemples de post-traitement de boîtes englobantes (élargissement-contraction). Les boîtes avant (resp. après) post-traitement sont représentées en rouge (resp. jaune).

plète s'appuyant sur une pyramide de descripteurs CNN et un ajustement du score de détection. Une analyse détaillée de ces résultats est fournie dans [TWDC17]. Ils montrent globalement qu'utiliser un pyramide de descripteurs CNN est plus performant que les traditionnels HOG, et que l'ajustement du score de détection permet d'améliorer de manière significative les performances de détection. Ainsi, notre méthode obtient les meilleurs résultats de ce comparatif pour les catégories «boat», «cat», «horse» et «train», et obtient une précision moyenne de 27,4%, ce qui est compétitif avec l'état de l'art.

2.3 Détection d'objets semi-supervisée basée sur le transfert de connaissances visuelles et sémantiques

Dans la section précédente, nous avons proposé une méthode de détection d'objets dans un cas faiblement supervisé, c'est à dire lorsque les annotations des catégories ne sont fournies qu'au niveau global de l'image. Nous nous intéressons ici au cas où l'annotation au niveau des objets est disponible pour certaines catégories.

method / class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	
our M-WDPM-HOG	34.1	41.5	15.2	10.0	8.8	36.5	40.8	31.5	4.6	23.1	
our M-WDPM-deep	38.2	38.4	17.5	15.8	9.5	38.1	39.4	32.0	3.5	26.4	
our M-WDPM-rescore	46.6	40.1	18.5	**18.1**	10.7	38.9	43.7	**38.9**	10.8	30.1	
Model Drift	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	
Multi-fold MIL	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0	
Min-Supervision	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	
Pattern Config	36.3	**47.6**	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5	
Posterior Reg.	42.2	43.9	23.1	9.2	**12.5**	**44.9**	45.1	24.9	8.3	24.0	
Convex Clustering	46.2	46.9	**24.1**	16.4	12.2	42.2	**47.1**	35.2	7.8	28.3	
LCL-pLSA	**48.8**	41.0	23.6	12.1	11.1	42.7	40.9	35.5	**11.1**	**36.6**	
DPM 5.0†	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	
DP-DPM conv5†	42.3	65.1	32.2	24.4	36.7	56.8	55.7	38.0	28.2	47.3	
R-CNN†	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	
method / class	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
our M-WDPM-HOG	9.4	24.2	29.8	42.5	9.1	14.5	18.3	11.2	32.1	14.3	22.6
our M-WDPM-deep	11.2	26.1	33.1	43.7	8.8	16.7	20.8	14.5	33.5	18.0	24.3
our M-WDPM-rescore	16.3	26.9	**37.4**	42.1	12.9	18.9	22.5	16.2	**38.1**	19.6	27.4
Model Drift	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0	13.9
Multi-fold MIL	4.9	14.1	31.9	41.9	**19.3**	11.1	**27.6**	12.1	31.0	40.6	22.4
Min-Supervision	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Pattern Config	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Posterior Reg.	13.9	18.6	31.6	43.6	7.6	**20.9**	26.6	20.6	35.9	29.6	26.4
Convex Clustering	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
LCL-pLSA	**18.4**	**35.3**	34.8	**51.3**	17.2	17.4	26.8	**32.8**	35.1	**45.6**	**30.9**
DPM 5.0†	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DP-DPM conv5†	37.1	39.2	61.0	56.4	52.2	26.6	47.0	35.0	51.2	56.1	44.4
R-CNN†	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5

FIGURE 2.5 – Comparaison de méthodes de détection d'objets faiblements supervisées sur PASCAL VOC 2007 en terme de précision moyenne (en %) sur l'ensemble de test. († méthodes supervisées utilisant des annotations au niveau des objets.). Références : Model Drift [ST11], Multi-fold MIL [CVS14], Min-Supervision [SGJ+14], Pattern Config [SLJD14], Posterior Reg. [BPT14], Convex Clustering [BPT15], LCL-pLSA [WHR+15], DPM 5.0 [FGMR10], DP-DPM conv5 [GIDM15], R-CNN [GDDM14].

Nous disposons donc de catégories «entièrement annotées» et de catégories «faiblement annotées». Pour les catégories «entièrement annotées», un nombre important d'images d'entraînement disposant d'annotation des labels au niveau des images et également au niveau des boîtes englobantes sont disponibles pour l'apprentissage non seulement de classifieurs d'images (utilisant les labels au niveau des images), mais également de détecteurs d'objets (utilisant les labels au niveau des boîtes englobantes). Pour les catégories «faiblement annotées», de nombreuses images contenant les objets cibles sont disponibles mais l'annotation étant au niveau global de l'image, nous n'avons pas accès à la position exacte des objets.

Une approche naïve consisterait à appliquer un classifieur d'images directement sur des boîtes englobantes pour réaliser la détection d'objets. Or, cela mènerait à des performances très médiocres du fait des importantes différences de distribution statistique entre les données d'apprentissage (images globales comportant de nom-

breux éléments dont objets et arrière-plan) et les données de test (boîtes englobantes contenant un unique objet). Ce problème a été abordé dans [HGT+14] et la stratégie proposée, LSDA pour «Large Scale Detection through Adaptation», consiste à apprendre une transformation entre des classifieurs CNN et des détecteurs d'objets pour des catégories avec à la fois des annotations au niveau image et au niveau objets (catégories «fortes»), puis d'appliquer cette transformation pour adapter des classifieurs d'images en détecteurs d'objets pour des catégories avec uniquement les annotations au niveau image (catégories «faibles»). Cela implique de transférer les connaissances de différences entre les classifieurs et détecteurs pour des catégories fortes aux classifieurs pour des catégories faibles et proches visuellement, afin d'obtenir les détecteurs pour ces catégories faibles. Cette problématique est illustrée dans la Figure 2.6.

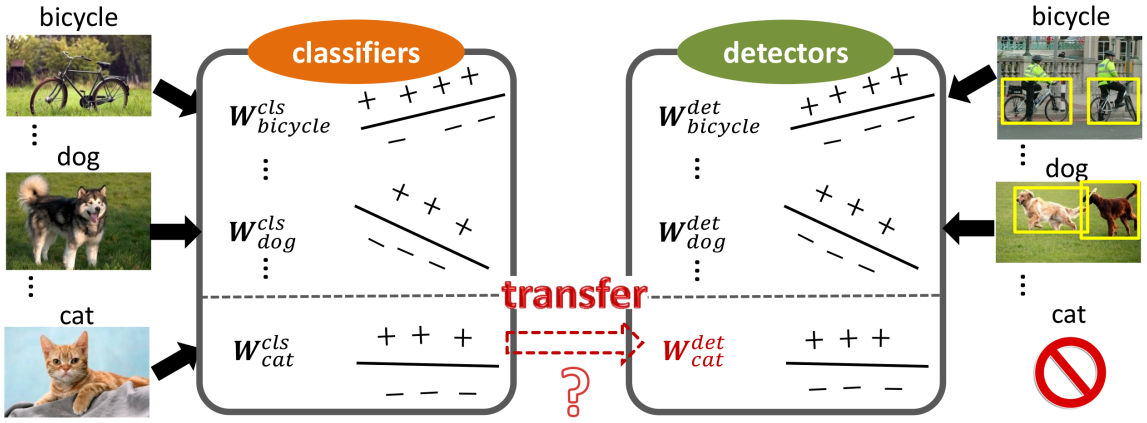


FIGURE 2.6 – Considérons un jeu de données contenant des labels au niveau des images pour toutes les catégories, et des annotations au niveau des objets (des boîtes englobantes) seulement pour quelques catégories (catégories fortement annotées) (cas de «bicyle» et «dog»). Comment alors transformer un classifieur d'images en un détecteur d'objets pour permettre la détection des objets des catégories faiblement annotées (exemple de «cat») ?

Dans [HGT+14], afin d'identifier les catégories fortes similaires à la catégorie faible pour laquelle une transformation d'un classifieur d'images en détecteurs d'objets doit être opérée, une mesure de similarité entre catégories est utilisée, reposant sur une distance dans l'espace des poids de la couche $fc8$ du réseau CNN de type Alex-Net [KSH12].

Notre contribution dans ce contexte est d'incorporer des connaissances sur les similarités entre objets d'un point de vue visuel mais également sémantique dans la modélisation des différences entre classifieurs et détecteurs pour les catégories fortes, et ensuite transférer ces connaissances pour adapter un classifieur en un détecteur

pour les catégories faibles. Notre démarche est ainsi motivée par les observations suivantes : (1) les différences entre classifieurs et détecteurs sont spécifiques à chaque catégorie [GDDM14, KSH12] ; (2) les catégories similaires visuellement et sémantiquement sont susceptibles de posséder d'avantage de propriétés communes transférables que des catégories éloignées visuellement et sémantiquement ; (3) la similarité visuelle et la liaison sémantique sont fortement corrélées, en particulier lorsqu'elles sont mesurées entre objets extraits des images (en éliminant l'arrière-plan) [DF11]. En effet, intuitivement, il serait plus souhaitable d'adapter un classifieur «chat» en détecteur «chat» en utilisant les différences spécifiques aux catégories entre le classifieur et le détecteur «chien», plutôt que «violon» ou «fraise». Cette procédure est illustrée Figure 2.7.

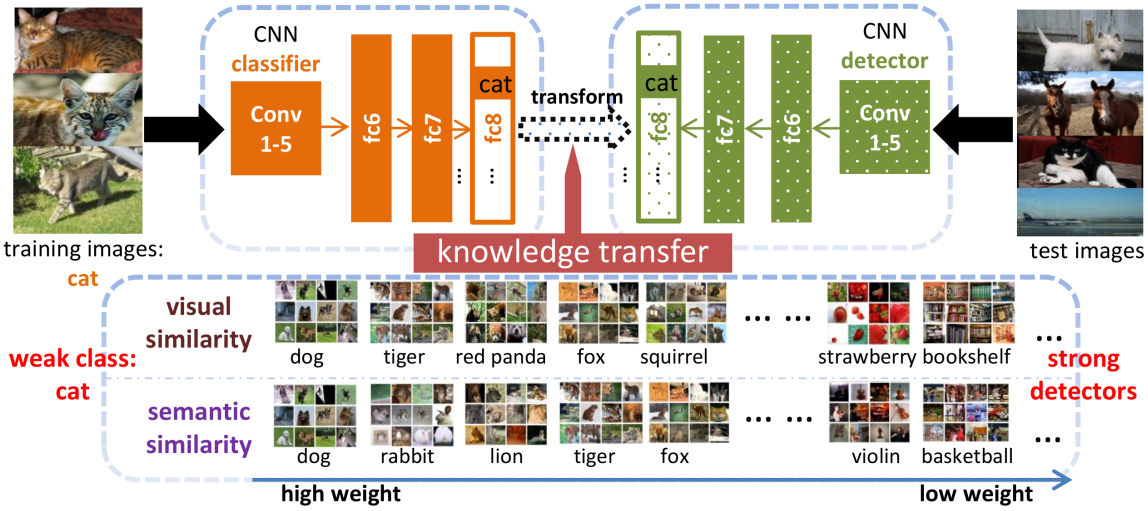


FIGURE 2.7 – Illustration de notre modèle de transfert de connaissances basé sur la similarité. Pour adapter un classifieur d'images (en haut à gauche) d'une catégorie «faiblement annotée» (pas d'annotation au niveau des boîtes englobantes) en un détecteur d'objets (en haut à droite), nous transférons les connaissances des différences entre classifieurs et détecteurs pour des catégories «entièrement annotées» (avec des annotations au niveau de l'image et également au niveau des boîtes englobantes, en bas de la figure) en favorisant les catégories les plus semblables à la catégorie cible. Par exemple, pour produire un détecteur «cat», le transfert d'information de «dog» et «tiger» sera privilégié, plutôt que celui de «basketball» ou «bookshelf».

Comme mentionné précédemment, l'approche que nous proposons s'appuie sur la méthode LSDA [HGT+14]. Nous allons donc rappeler son principe avant d'exposer nos contributions.

Principe de LSDA

Soit D le jeu de données des K catégories à détecter. Les annotations à la fois au niveau image et boîtes englobantes sont disponibles uniquement pour un ensemble m de catégories «entièrement annotées» ($m \ll K$) noté B. Pour l'ensemble A des autres catégories, celles «faiblement annotées», seules les annotations au niveau global de l'image sont disponibles. Ainsi, un ensemble de K classifieurs d'images peut être entraîné sur l'ensemble du jeu de données D ($D = A \cup B$), mais seulement m détecteurs d'objets (de B) peuvent être entraînés à partir des annotations de boîtes englobantes. L'algorithme LSDA apprend à convertir $(K - m)$ classifieurs d'images (de A) en détecteurs d'objets des catégories correspondantes par les étapes suivantes :

1. Pré-entraînement : Tout d'abord, un réseau de neurones convolutif CNN de type AlexNet [KSH12] avec 8 couches (5 de convolution et 3 entièrement connectées (fc pour «fully connected»)) est pré-entraîné sur le jeu de données de classification d'images ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [RDS+15], qui contient 1,2 million d'images de 1000 catégories.

2. Ajustement de l'entraînement pour la classification : La couche de sortie (1000 classifieurs linéaires) du CNN pré-entraîné est remplacée par une couche de dimension K (K classifieurs linéaires). Les poids des connections des neurones de cette couche sont initialisés aléatoirement et l'entraînement de l'ensemble du CNN est ensuite poursuivi sur le jeu de données D (fine tuning). Cela produit un réseau de classification pouvant classer K catégories en donnant en entrée une image ou une région d'une image.

3. Adaptation commune aux catégories : L'entraînement du réseau classifieur est ensuite poursuivi (toujours fine tuning) sur les boîtes englobantes de B en utilisant l'approche R-CNN [GDDM14]. Cela permet de transformer le classifieur en détecteur. Les paramètres de la couche $fc8$ sont spécifiques aux catégories, avec 4097 valeurs de poids pour chaque catégorie, alors que les paramètres des couches 1 à 7 sont communs à toutes les catégories. Dans la mesure où les détecteurs d'objets ne peuvent pas être entraînés sur A (pas d'annotation au niveau des boîtes englobantes), la couche de sortie spécifique aux catégories $fc8_A$ reste inchangée. La matrice des différences de $fc8_B$ après ce nouvel entraînement est noté Δ_B.

4. Adaptation spécifique aux catégories : Finalement, chaque classifieur pour les catégories $j \in A$ est adapté en détecteur correspondant par l'apprentissage d'une transformation des paramètres du modèle spécifique aux catégories. Ceci est basé sur l'hypothèse que la différence entre un classifieur et un détecteur pour une

catégorie donnée a une corrélation positive avec celles de catégories similaires (ou proches). Cette transformation est calculée en ajoutant un vecteur de biais aux poids de $fc8_A$. Ce vecteur de biais pour la catégorie j est obtenu en moyennant les changements de poids entre le classifieur et le détecteur des k catégories les plus proches dans l'ensemble B.

$$\forall j \in A : \overrightarrow{w_j^d} = \overrightarrow{w_j^c} + \frac{1}{k} \sum_{i=1}^{k} \Delta_{B_i^j} \tag{2.5}$$

où $\Delta_{B_i^j}$ correspond aux différences de poids de la couche $fc8$ de la $i^{ème}$ catégorie voisine dans l'ensemble B pour la catégorie $j \in A$. $\overrightarrow{w^c}$ et $\overrightarrow{w^d}$ sont respectivement les poids de la couche $fc8$ pour le classifieur ajusté par finetuning et pour le détecteur correspondant. Les catégories les plus proches sont définies comme celles avec la norme L_2 la plus proche (distance euclidienne) des poids $fc8$ dans l'ensemble B.

Hoffman et al. [HGT+14] ont montré que le modèle adapté par LSDA permet d'augmenter les performances de 50% en terme de précision moyenne (mAP) pour la détection par rapport à l'utilisation d'un classifieur appliqué directement à une fenêtre englobante pour les catégories faiblement annotées sur le jeu de données de détection ILSVRC2013 (de 10,31% à 16,15%). Ils ont également montré que l'adaptation spécifique aux catégories ($4^{ème}$ étape de LSDA) contribue relativement moins à l'amélioration des performances (16,15% contre 15,85% dans cette étape) que l'adaptation commune aux catégories (ajustement des poids des couches 1 à 7), qui serait donc la phase la plus importante.

Cependant, nos expérimentations ont montré qu'en ajustant de manière appropriée cette couche spécifique $fc8$, un gain de performance important peut être obtenu. C'est pourquoi nous avons proposé une stratégie permettant d'adapter cette couche en utilisant des connaissances sur les similarités entres catégories d'objets.

Transfert de connaissances par similarité visuelle

Intuitivement, le détecteur d'objet d'une catégorie devrait être plus semblable à celui de catégories visuellement similaires qu'à ceux de catégories visuellement différentes. Par exemple, un détecteur de chat devrait mieux approximer un détecteur de chien qu'un détecteur de fraise, puisque chat et chien sont tous deux des mammifères partageant des propriétés communes en terme de forme (une tête, quatre jambes, deux oreilles, deux yeux, une queue, ...) et de texture (tous les deux ont une fourrure). Ainsi, étant donné un jeu de données entièrement annoté B et un jeu de données faiblement annoté A, notre objectif est de modéliser la similarité

visuelle entre chaque catégorie $j \in A$ et toutes les autres catégories dans B, puis de transférer cette connaissance pour transformer des classifieurs en détecteurs pour A.

Les mesures de similarité visuelle sont souvent obtenues en calculant des distances entre des distributions de descripteurs telles que les sorties des couches $fc6$ ou $fc7$ d'un CNN, ou dans le cas de LSDA, les paramètres de la couche $fc8$. Nous proposons une approche différente qui consiste à propager une image à travers tout le réseau ajusté (étape 2 de LSDA) afin d'obtenir un vecteur des scores de classification à K dimensions. Ce vecteur représente les probabilités d'une image d'appartenance aux K catégories. Ainsi, pour chaque image positive d'une catégorie $j \in A$, il est possible d'accumuler les scores pour chaque dimension, sur un jeu de données de validation balancé (nombre égal d'images par catégorie). Nous considérons alors que les scores accumulés normalisés (valeurs entres 0 et 1) correspondent aux similarités entre la catégorie j et toutes les autres catégories : plus le score est élevé, et plus la catégorie correspondante est visuellement proche de la catégorie j. Cette hypothèse est supportée par l'analyse de CNNs profonds [AGM14, ZF14] : les réseaux CNN sont susceptibles de confondre des catégories visuellement similaires, sur lesquelles ils auront des scores de prédiction plus élevés.

La similarité visuelle, notée s_v, entre une catégorie faiblement annotée $i \in A$ et une catégorie entièrement annotée $j \in B$ est définie par :

$$s_v(j,i) \propto \frac{1}{N} \sum_{n=1}^{N} CNN_{softmax}(I_n)_i \tag{2.6}$$

où I_n est une image positive de la catégorie j de l'ensemble de validation de A, N est le nombre d'images positives pour cette catégorie, et $CNN_{softmax}(I_n)_i$ est la valeur de la $i^{ème}$ sortie de la couche softmax du CNN pour l'image I_n, c'est à dire la probabilité pour I_n d'appartenir à la catégorie $i \in B$. $s_v(j,i) \in [0,1]$ est donc le degré de similarité après normalisation par rapport à toutes les catégories de B.

En utilisant l'équation (2.5), il est possible de transférer les paramètres du modèle basé sur les k catégories les plus proches d'une catégorie donnée, sélectionnées par le critère de similarité visuel exprimé dans l'équation (2.6). Une alternative à l'équation (2.5) consiste à considérer un schéma de pondération des voisins les plus proches, dans lequel les poids attribués aux différentes catégories reflètent la similarité visuelle de ces catégories avec la catégorie cible. En effet, intuitivement, différentes catégories auront différents degrés de similarité avec une catégorie particulière, et certaines catégories auront très peu (ou beaucoup) de catégories visuellement similaires. Nous avons donc modifié l'équation (2.5) et défini la transformation suivante, basée sur la similarité visuelle reposant sur le schéma de pondération des

voisins les plus proches :

$$\forall j \in A : \overrightarrow{w^d_{j\,v}} = \overrightarrow{w^c_j} + \sum_{i=1}^{m} s_v(j,i)\Delta_{B^j_i} \tag{2.7}$$

On peut noter que l'équation (2.5) est un cas particulier de l'équation (2.7) où $m = k$ et $s_v(j,i) = 1/k$.

Transfert de connaissances par similarité sémantique

Dans la lignée des travaux présentés dans [DF11, RW15, RSS+10], nous avons observé que la similarité visuelle est fortement corrélée à la liaison sémantique, en particulier lorsque l'on s'attache aux instances d'objets dans les images, ignorant l'arrière-plan. Ainsi, en s'appuyant sur ce fait, nous avons proposé de transférer des connaissances du domaine du langage naturel pour aider à améliorer la détection d'objets semi-supervisée.

Nous avons pour cela utilisé la méthode de «plongement lexical» («word embeddings») [MSC+13, PSM14] pour représenter chaque catégorie et mesurer la similarité sémantique entre catégories. En effet, cette méthode permet de représenter chaque terme par un vecteur contextuel, et deux termes sémantiquement proches auront des vecteurs très similaires car ils apparaîtront généralement simultanément dans les mêmes contextes.

Ainsi, chacune des K catégories est représentée par un vecteur «word2vec» à 300 dimensions [MSC+13]. Comme chaque catégorie du jeu de données ImageNet [RDS+15] que nous utilisons dans nos expériences est un synset WordNet [Fel98], elles sont représentées comme la somme des vecteurs pour chaque terme dans son synset, normalisé au vecteur unitaire par sa norme L_2. Ainsi, la distance sémantique entre chaque catégorie $j \in A$ et $i \in B$ peut être obtenue par la norme L_2 $d_s(j,i)$ de chaque paire. La similarité sémantique $s_s(j,i)$ est donc inversement proportionnelle à $d_s(j,i)$.

Finalement, de manière semblable à la transformation exprimée dans l'équation (2.7), la transformation basée sur la similarité sémantique reposant sur le schéma de pondération des voisins les plus proches peut être donnée par :

$$\forall j \in A : \overrightarrow{w^d_{j\,s}} = \overrightarrow{w^c_j} + \sum_{i=1}^{m} s_s(j,i)\Delta_{B^j_i} \tag{2.8}$$

Modèle de transfert combiné

Nous avons proposé deux modèles différents de transfert de connaissances. Chacun d'eux peut être intégré dans la méthode LSDA indépendamment. De plus, comme nous considérons la similarité visuelle au niveau de l'image globale et la liaison sémantique au niveau des objets, ils peuvent être combinés simultanément pour fournir une information complémentaire. Nous avons utilisé une combinaison simple mais efficace pour constituer notre modèle final de transfert de connaissances. Il s'agit d'une combinaison linaire des similarités visuelles et sémantiques, selon l'équation (2.9) :

$$s = intersect[\alpha s_v + (1 - \alpha) s_s] \tag{2.9}$$

où $intersect[.]$ est une fonction sélectionnant les catégories co-occurentes parmi les catégories similaires visuellement et sémantiquement. α est un hyper-paramètre contrôlant l'influence relative des deux mesures de similarité.

Quelques exemples de similarités visuelles et sémantiques entre catégories sont données dans la Figure 2.8. On peut observer que la liaison sémantique est généralement corrélée à la similarité visuelle.

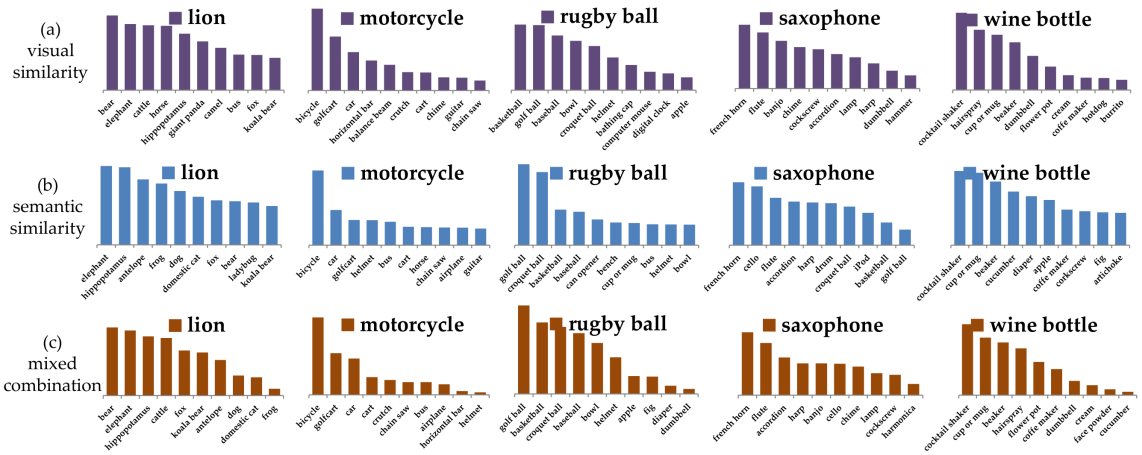


FIGURE 2.8 – Quelques exemples de similarité visuelle (a), de similarité sémantique (b) et de combinaison de similarités (c) entre une catégorie cible faiblement supervisée et les catégories source à partir desquelles la connaissance doit être transférée. Pour chaque catégorie cible, les 10 catégories source les plus proches sont présentées. L'amplitude de chaque colonne correspond au poids de la catégorie correspondante (degré de similarité s_v, s_s, s)

Transfert pour la régression des boîtes englobantes

Les fenêtres de détection générées par les modèles de détection basés région, tels que Selective Search [USGS13], sont généralement approximatives et il est souvent nécessaire d'améliorer leur précision de localisation par une méthode de post-traitement. Cela peut être fait par des méthodes de régression de boîtes englobantes, telles que celle présentée dans [GDDM14]. Cependant, ces méthodes nécessitent de disposer des annotations au niveau des objets, ce qui est un obstacle dans le cas de catégories faiblement supervisées. Ainsi, nous avons également proposé de transférer aux catégories faiblement annotées les connaissances de régresseurs spécifiques entraînés sur les catégories entièrement annotées, en se basant sur les mesures de similarité visuelles et sémantiques décrites précédemment.

Afin d'entraîner un régresseur pour une catégorie entièrement annotée, N paires d'apprentissage sont sélectionnées $\{(\overrightarrow{P^i}, \overrightarrow{G^i})\}_{i=1,\dots,N}$ où $\overrightarrow{P^i} = (P^i_x, P^i_y, P^i_w, P^i_h)$ est un vecteur indiquant les coordonnées du centre (P^i_x, P^i_y) de la proposition de fenêtre P^i ainsi que sa largeur et sa hauteur (P^i_w, P^i_h). $\overrightarrow{G^i} = (G^i_x, G^i_y, G^i_w, G^i_h)$ représente la boîte englobante correspondante réelle (vérité terrain). Dans la suite, nous omettons l'exposant i sauf quand nécessaire pour éviter toute confusion. L'objectif est alors d'apprendre une fonction de correspondance $f(P) = (f_x(P), f_y(P), f_w(P), f_h(P))$ faisant correspondre à chaque proposition de région P la fenêtre de vérité terrain G. Chaque fonction dans $f(P)$ est modélisée comme une fonction linéaire des descripteurs *pool*5 (la carte de descripteurs après le dernier bloc de convolution et pooling du réseau convolutif) : $f(P) = w_*^T F_5(P)$ où w_* est un vecteur de paramètres à apprendre et $F_5(P)$ est le descripteur *pool*5 de la proposition de région P. w_* peut être appris en optimisant la fonction objectif suivante :

$$w_* = \underset{\hat{w_*}}{\operatorname{argmin}} \sum_{i=1}^{N} (\hat{w_*}^T F_5(P^i) - t^i_*)^2 + \lambda_0 ||\hat{w_*}||^2 \tag{2.10}$$

où $t_* = (t_x, t_y, t_w, t_h)$ est la cible de la régression pour la paire d'entraînement (P, G) définie par :

$$t_x = (G_x - P_x)/P_w, \tag{2.11}$$

$$t_y = (G_y - P_y)/P_h, \tag{2.12}$$

$$t_w = log(G_w/P_w), \tag{2.13}$$

$$t_h = log(G_h/P_h). \tag{2.14}$$

Les deux premières équations spécifient une translation invariantes à l'échelle du centre de la boîte englobante, alors que les deux équations suivantes spécifient une translation dans l'espace logarithmique de la largeur et de la hauteur de la boîte englobante. Après avoir appris les paramètres de la fonction de transformation, une fenêtre de détection (proposition de région) P peut être transformée en une nouvelle prédiction $\hat{P} = (\hat{P}_x, \hat{P}_y, \hat{P}_w, \hat{P}_h)$ en appliquant :

$$\hat{P}_x = P_x + P_w f_w(P), \tag{2.15}$$

$$\hat{P}_y = P_y + P_h f_h(P), \tag{2.16}$$

$$\hat{P}_w = P_w exp(f_w(P)), \tag{2.17}$$

$$\hat{P}_h = P_h exp(f_h(P)). \tag{2.18}$$

La paire d'entraînement (P, G) est sélectionnée si la proposition P a une valeur élevée (au delà d'un seuil fixé) de la mesure IoU (Intersection over Union) avec la boîte englobante réelle G.

Pour une catégorie faiblement annotée j, la fonction de transformation ne peut pas être apprise explicitement en raison de l'absence d'annotation des boîtes englobantes. Cependant, il est toujours possible de transférer cette connaissance à partir de catégories similaires de l'ensemble entièrement annoté B :

$$\forall j \in A : w_j = \sum_{i=1}^{m} s_* w_i \tag{2.19}$$

où s_* indique l'une des mesures de similarité présentées précédemment.

Evaluation expérimentale

Nous avons évalué nos modèles de transfert de connaissances pour la détection d'objet semi-supervisée sur le jeu de données ILSVRC2013 couvrant 200 catégories d'objets. A des fins de comparaison, nous avons suivi tous les réglages expérimentaux décrits dans [HGT+14], en particulier, nous avons simulé avoir accès aux 200 catégories pour les annotations au niveau global des images, et seulement aux 100 premières catégories (par ordre alphabétique) pour les annotations des boîtes englobantes. Les résultats de détection sont donnés dans la Figure 2.9.

Dans la mesure où notre objectif est la détection des catégories faiblement annotées, nous nous concentrons principalement sur la deuxième colonne (précision moyenne sur l'ensemble A). Les lignes de 1 à 5 sont les résultats de référence obte-

Method	Number of Nearest Neighbors	mAP on $\mathcal{B}$: "Fully labeled" 100 Categories	mAP on $\mathcal{A}$: "Weakly labeled" 100 Categories	mAP on $\mathcal{D}$: All 200 Categories
Classification Network	-	12.63	10.31	11.90
LSDA (only class invariant adaptation)	-	27.81	15.85	21.83
LSDA (class invariant & specific adapt)	avg/weighted - 5	28.12 / –	15.97 / 16.12	22.05 / 22.12
	avg/weighted - 10	27.95 / –	16.15 / 16.28	22.05 / 22.12
	avg/weighted - 100	27.91 / –	15.96 / 16.33	21.94 / 22.12
Ours (visual transfer)	avg/weighted - 5	27.99 / –	17.42 / 17.59	22.71 / 22.79
	avg/weighted - 10	27.89 / –	17.62 / 18.41	22.76 / 23.15
	avg/weighted - 100	28.30 / –	17.38 / 19.02	22.84 / 23.66
Ours (semantic transfer)	avg/weighted - 5	28.01 / –	17.32 / 17.53	22.67 / 22.77
	avg/weighted - 10	28.00 / –	16.67 / 17.50	22.31 / 22.75
	avg/weighted - 100	28.14 / –	17.04 / 18.32	23.23 / 23.28
	Sparse rep. - $\leq$20	28.18	19.04	23.66
Ours (mixture transfer)	-	28.04	**20.03** ↑3.88	24.04
Ours (mixture transfer + BB reg.)	-	31.85	**21.88**	26.87
Oracle: Full Detection Network (no BB reg.)	-	29.72	26.25	28.00
Oracle: Full Detection Network (BB reg.)	-	32.17	29.46	30.82

FIGURE 2.9 – Résultats de détection en terme de précision moyenne (mAP) sur l'ensemble de validation de ILSVRC2013.

nus avec LSDA. La première ligne indique les résultats de détection en appliquant directement un classifieur d'images entraîné uniquement avec des données de classification, obtenant une précision moyenne de 10,31% sur les 100 catégories faiblement annotées. La dernière ligne indique les résultats d'un réseau de détection oracle considérant que l'annotation des boîtes englobantes est disponible pour les 200 catégories (apprentissage supervisé). Cela correspond à la marge supérieure accessible (26,25%). Nos modèles de transfert de connaissance visuel et sémantique se révèlent être plus performants que LSDA, et les meilleurs résultats sont obtenus avec notre modèle combinant les deux similarités visuelles et sémantiques avec la régression des boîtes englobantes (Bounding Boxes), prouvant que la prise en compte de ces deux connaissances joue un rôle essentiel dans l'amélioration du processus d'adaptation de classifieurs d'images en détecteurs d'objets. Quelques exemples de détections réussies sont donnés dans la Figure 2.10. Une analyse plus approfondie des résultats peut être trouvée dans [TWW+18].

2.4 Conclusion

Nous avons présenté dans ce chapitre nos contributions pour le problème de la détection d'objets dans les images, avec en particulier une approche adaptée à un apprentissage faiblement supervisé, et une seconde à un apprentissage semi-supervisé.

Le premier modèle, dans le cas faiblement supervisé, s'appuie sur l'approche

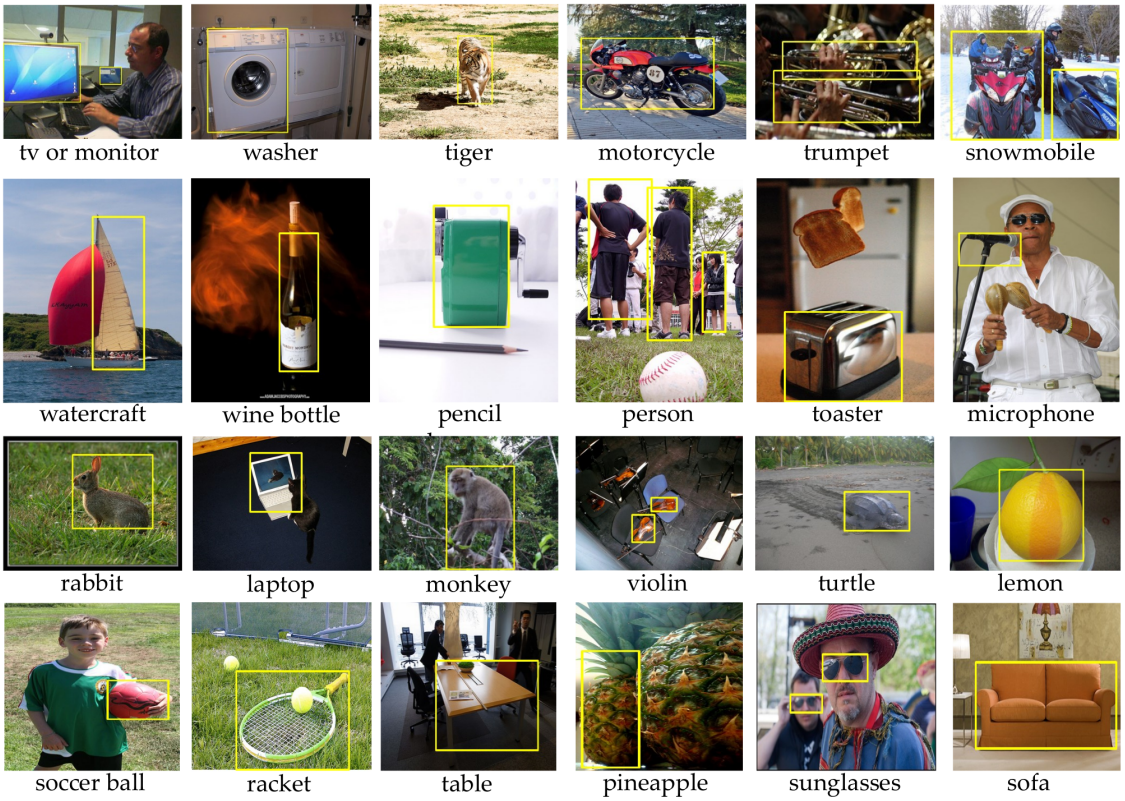


FIGURE 2.10 – Quelques exemples de détection réussie de notre modèle de transfert de connaissances combinant les similarités visuelles et sémantiques sur des images du jeu de données ILSVRC2013.

DPM «Deformable Part-based Models» et en propose plusieurs améliorations : une nouvelle méthode basée sur la notion «objectness» et sur la saillance visuelle pour sélectionner de manière adaptative un ensemble de fenêtres candidates susceptibles de contenir les objets à détecter dans l'image, ainsi qu'un procédé d'apprentissage de classes latentes pour classer une fenêtre soit en objet cible, soit en catégorie non-cible. Une méthode de post-traitement flexible permettant d'améliorer la position et la taille des boîtes englobantes fournies en sortie du DPM a également était introduite. Les expérimentations conduites sur le jeu de données PASCAL VOC 2007 ont montré que notre approche est compétitive avec celles de l'état de l'art.

Pour le cas semi-supervisé, notre stratégie s'appuie sur la transformation de classifieurs d'images en détecteurs d'objets, en transférant des connaissances sur les similarités visuelles et sémantiques entre les classes d'objets. Nos expérimentations de diverses architectures basées sur des réseaux convolutifs sur le jeu de données ILSVRC2013 ont démontré qu'à la fois la similarité visuelle et la similarité sémantique jouent un rôle essentiel dans le processus de transformation, suggérant que la connaissance inhérente à ces deux domaines est complémentaire, ce qui a permis à

notre approche d'obtenir de très bonnes performances.

Chapitre 3

Prédiction de l'impact émotionnel des vidéos

3.1 Introduction

Une première application de nos travaux portant sur l'analyse visuelle concerne l'informatique affective, et plus précisément la reconnaissance de l'émotion suscitée par les vidéos, c'est à dire l'émotion ressentie lors du visionnage d'une vidéo. Ceci a de nombreuses applications telles que la distribution de contenus personnalisés basés sur l'émotion [Han06] ou encore l'indexation et le résumé de vidéos [GLFM15, GLFM16]. Alors que d'importants progrès ont été réalisés en vision par ordinateur et apprentissage automatique notamment pour la compréhension de scènes visuelles, une étape suivante consiste à modéliser et reconnaître les concepts affectifs, le but étant de doter les ordinateurs de capacités de perception semblables à celles des humains. Ceci représente un challenge très relevé, notamment en raison de la complexité et de la nature subjective des émotions.

Dans ce cadre, nous avons proposé deux contributions. La première concerne la création d'une base volumineuse et fiable de vidéos annotées selon l'émotion pouvant être accessible aux chercheurs de la communauté afin d'élaborer et d'améliorer des modèles de prédiction de l'émotion induite. La deuxième contribution concerne justement l'élaboration de modèles, notamment basés sur l'apprentissage profond. Ces travaux ont été réalisés avec le doctorant Yoann Baveye dans le cadre d'une collaboration avec l'entreprise Technicolor et ont été publiés dans [R7, B1, B2, C5, C8, C9, W1, W2, A1] (numérotation correspondant aux publications listées dans le rapport d'activité en fin de document).

3.2 LIRIS-ACCEDE : une plateforme de données pour l'analyse du contenu émotionnel de vidéos

L'apprentissage efficace de modèles de prédiction et leur évaluation nécessite un jeu de données avec un contenu très riche et varié, associé à des annotations correspondant à la vérité terrain. Or ces données sont souvent très difficiles à collecter, ce qui est encore plus vrai lorsqu'il s'agit des émotions. Pour résoudre ce problème, nous avons créé la plateforme de données LIRIS-ACCEDE [1]. Contrairement à la plupart des jeux de données existants qui contiennent en général relativement peu de données et avec un accès limité en raison de contraintes de copyright, LIRIS-ACCEDE contient un grand nombre de vidéos variées annotées selon l'émotion. Toutes les vidéos sont sous licence "Creative Commons", ce qui leur permet donc d'être librement diffusables. La plateforme de données LIRIS-ACCEDE (vidéos, annotations, descripteurs et protocoles) est donc librement disponible, et est actuellement composé de six collections décrites dans les sections suivantes.

La qualité de cette plateforme a été reconnue d'une part par plusieurs publications dans les conférences et journaux du domaine de l'informatique affective et d'autre part par son adoption comme données d'apprentissage et de test pour les tâches "Affective Impact of Movies" à MediaEval 2015 [SBW+15], et "Emotional Impact of Movies" à MediaEval 2016 [DCB+16], 2017 [DHC+17] et 2018 [DHC+18], MediaEval proposant chaque année des tâches pour l'évaluation scientifique de méthodes destinées à l'accès, l'indexation et l'interprétation de données multimédia [LSG+17]. Le nombre de téléchargements de LIRIS-ACCEDE est actuellement de 439 (janvier 2020).

Représentation des émotions

Comme indiqué précédemment, nous nous intéressons à la reconnaissance automatique des émotions suscitées par les vidéos. Dans ce contexte, trois types d'émotions peuvent être considérées : l'émotion prévue, induite et attendue [Han06]. L'émotion prévue («intended emotion») est celle que le cinéaste veut induire chez les spectateurs. L'émotion induite («induced emotion») est celle qu'un spectateur ressent lors du visionnage d'un film. Finalement, l'émotion attendue («expected emotion») est celle que la majorité de l'audience ressent en réponse au même contenu.

1. http ://liris-accede.ec-lyon.fr/

Alors que l'émotion induite est subjective et dépendante du contexte, l'émotion attendue peut être considérée comme objective puisqu'elle reflète la réponse plus ou moins unanime d'une audience à un stimulus donné [Han06]. C'est ce type d'émotion que nous considèrerons par la suite.

Diverses représentation des émotions ont été proposées dans la littérature, en particulier les représentations catégorielles et dimensionnelles.

Les représentations catégorielles des émotions sont très naturelles puisqu'elles remontent aux origines du langage et l'émergence de mots et expressions pour désigner des états émotionnels clairement séparables. De nombreuses catégorisations discrètes des émotions ont été proposées, telles que les six émotions basiques universelles proposées par Ekman dans [Ekm99], ou les huit émotions primaires définies par Plutchick [PLU80]. Ekman considère les émotions comme étant des états discrets associés aux expressions faciales. Il a ainsi postulé que leur nombre est fixe et qu'il s'agit des émotions basiques peur, colère, tristesse, dégoût, joie et surprise. Cette liste est sans doute la plus utilisée dans la littérature. Plutchick quant à lui a suggéré huit émotions : colère, peur, tristesse, dégoût, surprise, anticipation, confiance et joie. Selon lui, ces émotions primaires sont des primitives biologiques et ont évolué afin d'augmenter l'aptitude de l'animal à la reproduction. Cette représentation catégorielle n'est pas très adaptée à notre cas car le nombre d'émotions est trop petit par rapport à la diversité des émotions ressenties par les spectateurs de films. De plus, si le nombre de classes venait à augmenter, des ambiguïtés apparaîtraient en raison des difficultés de langage ou d'interprétations personnelles.

Les représentations dimensionnelles quant à elles correspondent à une modélisation des émotions comme des points dans un espace continu à n dimensions. Le modèle le plus célèbre est l'espace valence-arousal-dominance, connu également comme pleasure-arousal-dominance (PAD) proposé par Russel et Mehrabian [RM77] et largement utilisé dans les travaux sur la compréhension des émotions. Dans cet espace, chaque émotion subjective peut être décrite par sa position selon les dimensions de valence, arousal et dominance. La plage de valeur de la valence s'étend de négative (par exemple triste, déçu) à positive (joyeux, exalté), alors que l'arousal s'étend d'inactive (fatigué, songeur) à active (alarmé, en colère) et la dominance s'étend de dominé (ennuyé, triste) à «en contrôle» (excité, enchanté). Etant donné la difficulté à identifier de manière cohérente une troisième dimension (telle que la dominance), qui diffère de l'arousal, de nombreuses études (dont la nôtre) se limitent à la valence et l'arousal. En effet, en particulier pour les émotions suscitées par les vidéos, la valence et l'arousal totalisent la quasi-totalité de la variance [GCL89, LBC99]. C'est donc cette représentation des émotions que nous avons choisi de considérer pour la

plateforme de données LIRIS-ACCEDE.

Sélection des films

Un des objectifs principaux de LIRIS-ACCEDE était qu'elle devait être librement accessible par la communauté de recherche. C'est la raison pour laquelle nous avons sélectionné 160 films sous licence «Creative Commons» afin de la constituer. Cette licence nous permet en effet de partager publiquement la base sans problème de copyright.

La plupart des 160 films sont issus des plateformes de vidéos VODO et Vimeo. Les films disponibles sur ces plateformes sont créés par des cinéastes avec une excellente expertise technique et nombre d'entre eux ont été projetés lors de festivals de cinéma. Ainsi, parmi les 160, 40 sont des films de qualité élevée et 120 sont des films courts. La durée totale est de 73 heures, 41 minutes et 7 secondes. Les 9 genres représentatifs de ces films sont Comédie, Animation, Action, Aventure, Thriller, Documentaire, Romance, Drame et Horreur. En observant la distribution normalisée des films par genre dans LIRIS-ACCEDE, on constate qu'elle est similaire aux distributions des films référencés par IMDB et ScreenRush, comme l'indique la Figure 3.1. Les films dans LIRIS-ACCEDE sont donc représentatifs des films actuels. La langue est principalement l'anglais.

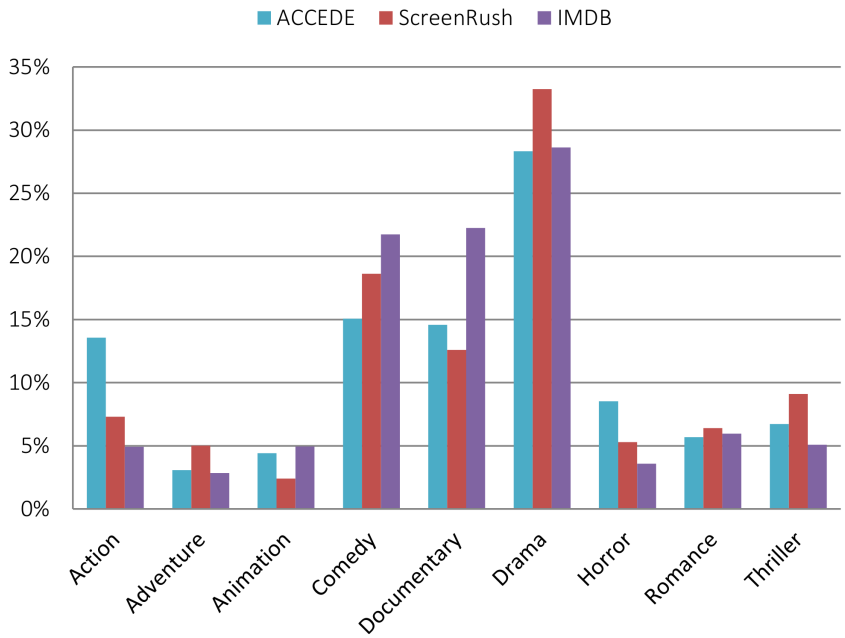


FIGURE 3.1 – Distribution normalisée des genres des films inclus dans LIRIS-ACCEDE et référencés par ScreenRush et IMDB.

Collection «Discrete LIRIS-ACCEDE»

Cette collection est constituée de 9800 clips vidéo extraits des 160 films en utilisant une méthode robuste de détection de changements de plans et de fondus implémentée à partir des algorithmes décrits dans [Lie98]. Puisque chaque clip commence et se termine par un changement de plan, il est très vraisemblable qu'ils seront perçus par les spectateurs comme sémantiquement cohérents. Les 9800 clips durent entre 8 et 12 secondes, et la durée totale est de 26 heures, 57 minutes et 8 secondes. Même si la résolution temporelle, ou granularité, des émotions fait encore débat, la plupart des psychologues s'accordent à dire qu'elles sont une partie d'un processus complexe mais très rapide, de l'ordre de quelques secondes [RRG07]. Ainsi, la durée des clips de LIRIS-ACCEDE est suffisamment longue pour qu'ils soient consistants, permettant au spectateur de ressentir une émotion, et suffisamment courte pour éviter que plusieurs émotions se succèdent [GL95]. Nous avons ainsi obtenu une grande variété de clips vidéo reflétant la variété des films sélectionnés. Ainsi, ces clips contiennent des scènes de violence, de sexe, de crimes mais également des paysages, des interviews et des scènes positives de la vie de tous les jours. Cela fait de cette collection la base de vidéos annotées selon l'émotion ayant la plus grande diversité de contextes.

Le processus d'annotation a pour objectif de trier les 9800 clips vidéo indépendamment selon les axes de valence et d'arousal (émotions attendues). Nous avons adopté une approche d'annotation par comparaisons qui est beaucoup plus fiable et robuste qu'une approche par évaluation, en particulier dans un contexte non contrôlé. En effet, demander à un annotateur de comparer l'émotion suscitée par le visionnage de deux vidéos est beaucoup plus aisé que lui demander une valeur absolue caractérisant l'émotion ressentie lors du visionnage d'une vidéo, ce qui est beaucoup plus subjectif. Ainsi, en choisissant des comparaisons deux à deux plutôt qu'une évaluation, les chances d'obtenir des annotations consistantes sont beaucoup plus élevées car les annotateurs auront plus tendance à s'accorder en décrivant des émotions en des termes relatifs plutôt qu'absolus [MN13]. Yang et Chen ont également montré dans [YC11] que des approches de comparaisons deux à deux augmentent la fiabilité des annotations comparées à des approches d'évaluation, notamment car le processus d'annotation est simplifié. De plus, cette simplification rend la tâche d'annotation plus attractive et intéressante pour les annotateurs.

Pour l'annotation de cette collection, qui suit donc une stratégie de comparaisons deux à deux, les annotateurs se verront présenter des paires de clips vidéo et ils devront à chaque fois sélectionner le clip suscitant la plus forte émotion en terme de valence et d'arousal.

Comme l'objectif est d'obtenir les 9800 clips vidéo ordonnés selon un axe de valence et un axe d'arousal, le nombre de paires à générer serait de $9800^2 = 96040000$, ce qui est bien entendu infaisable, d'autant plus que pour chaque paire trois jugements par des annotateurs différents doivent être fournis. Nous avons donc utilisé l'algorithme de tri quicksort de complexité moyenne $O(n \log n)$ (en comparaison à $O(n^2)$) pour générer les paires en fonction des résultats d'annotation à chaque itération. Le nombre de paires à annoter restant très élevé, nous avons réalisé les annotations par crowdsourcing [LMMX14] en utilisant le service Crowdflower devenu Figure Eight [2]).

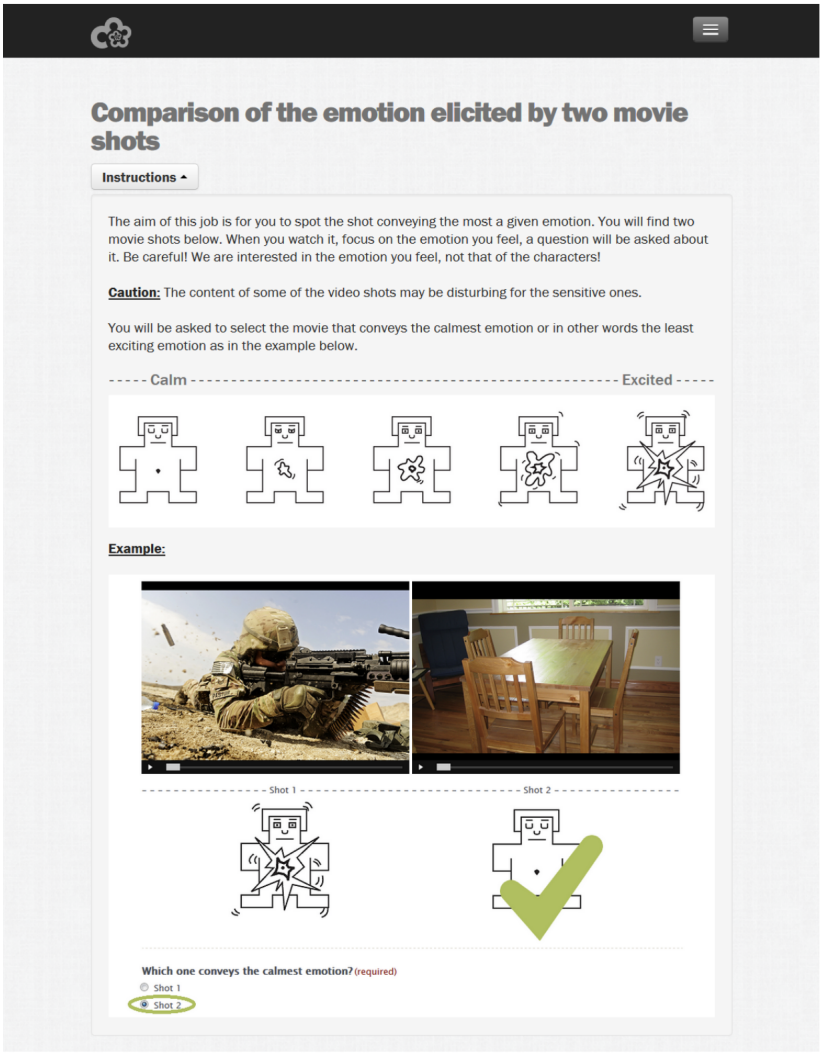


FIGURE 3.2 – Interface présentée aux annotateurs sur la plateforme de crowdsourcing pour l'annotation de l'axe arousal.

Toujours dans le but de simplifier le processus d'annotation et d'éviter une surcharge cognitive, l'annotation de la valence et de l'arousal ont été réalisées séparé-

2. www.figure-eight.com

ment.

La Figure 3.2 montre l'interface présentée aux annotateurs sur la plateforme de crowdsourcing pour l'annotation de l'axe arousal.

Pour une paire de clips vidéo donnée, les annotateurs doivent sélectionner le clip suscitant l'émotion la plus positive (pour la valence) ou l'émotion la plus calme (pour l'arousal). Les termes valence et arousal n'ont pas été utilisés afin d'éviter un incompréhension de la part des annotateurs. Il leur est demandé de se concentrer sur l'émotion qu'ils ressentent lors du visionage des clips vidéo (émotion induite) et non pas sur celle des acteurs. Pour faciliter la compréhension des annotateurs, l'échelle «Self-Assessment Manikin» [BL94] a été utilisée. Il s'agit d'un système de pictogrammes très intéressant car très compréhensible pour représenter les axes de valence, arousal et dominance. Afin de s'assurer de la fiabilité des annotateurs, un système de vérification a été utilisé pour détecter les annotateurs ne répondant pas sérieusement. Ceci permet de ne pas prendre en compte leurs annotations. Chaque paire de clips vidéo a été présentée aux annotateurs jusqu'à ce que trois annotation aient été collectées. Cela nous semblait un bon compromis entre le coût financier de l'annotation et la fiabilité de l'expérimentation.

Ce procédé nous permet donc finalement d'obtenir pour chaque clip vidéo deux valeurs discrètes variant de 0 à 9799 représentant son rang de valence et d'arousal.

Afin de nous assurer de la cohérence des annotations, nous avons étudié la fiabilité inter-annotateurs. Cela donne en effet une indication sur la capacité d'annotateurs indépendant à participer à une expérience et à atteindre la même conclusion malgré la subjectivité de la tâche. Plusieurs mesures de fiabilité inter-annotateurs ont été proposées dans la littérature, telles que «percent agreement», «Fleiss's kappa» [Fle71], «Krippendorff's alpha» [Kri70] ou encore «Randolph's kappa» [Ran10]. L'idée générale de ces mesures est d'indiquer si les annotateurs ont tendance à s'accorder sur leurs réponses. Les résultats sont présentés dans la Figure 3.3. Le «percent agreement» indique que les annotateurs ont été d'accord sur 86,2% et 83,5% des comparaisons. Pour la mesure Randolph's kappa, Landis et Koch [LK77] suggèrent qu'un score de 0,375 indique un bon accord et un score de 0,452 correspond à un accord modéré. Ainsi, ces résultats montrent que les annotateurs ont bien compris la tâche et ont fourni des annotations cohérentes malgré la subjectivité des deux expériences d'annotation. Une analyse plus complète peut être trouvée dans [BDCC15].

Measure	Arousal	Valence
Percent agreement	0.862	0.835
Fleiss' κ	0.190	0.179
Krippendorff's α	0.191	0.180
Randolph's κ	0.452	0.375

FIGURE 3.3 – Fiabilité inter-annotateurs.

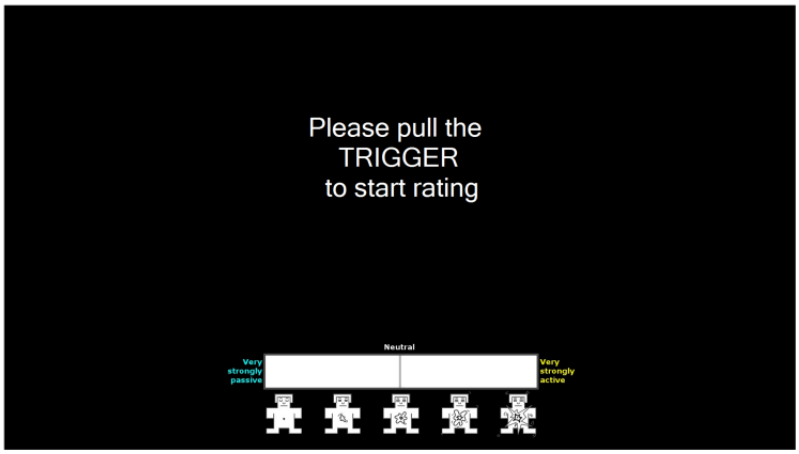
Collection «Continuous LIRIS-ACCEDE»

La collection «Discrete LIRIS-ACCEDE» présentée dans la section précédente propose donc 9800 clips vidéo extraits des 160 films. Ces clips, d'une durée variant entre 8 et 12 secondes, ont été annotés globalement puisque chacun d'eux est caractérisé par un rang de valence et un rang d'arousal. Afin de permettre des études plus approfondies sur les dépendances temporelles des émotions (puisqu'une émotion ressentie à un moment peut influencer l'émotion ressentie à un moment ultérieur), des films plus longs sont considérés dans cette collection. Pour cela, 30 films parmi les 160 ont été sélectionnés de manière à ce que leur genre, contenu, langue et durée soit suffisamment variés pour être représentatifs des données de la collection discrète. Ces vidéos sont d'une durée variant de 117 secondes à 4566 secondes (884,2 secondes en moyenne, avec un écart-type de 766,7 secondes). La durée totale est de 7 heures, 22 minutes et 5 secondes.

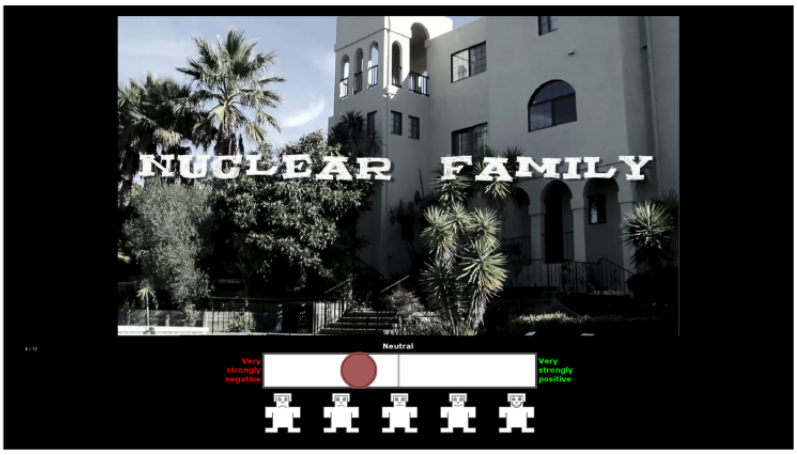
L'objectif de l'annotation de ces films est d'obtenir une auto-évaluation continue de la valence et d'arousal ressentie par les spectateurs regardant ces films.

Afin de collecter ces annotations continues, nous avons modifié le programme GTrace développé à l'origine par Cowie et al. [CSD+13] afin de l'adapter à nos besoins. Ainsi, les annotateurs vont pouvoir visualiser les films en plein écran, entendre le son par un casque et utiliser un joystick avec des mouvements gauche-droite pour indiquer le niveau d'émotion ressenti à chaque instant. Une illustration de l'interface est présentée dans la Figure 3.4

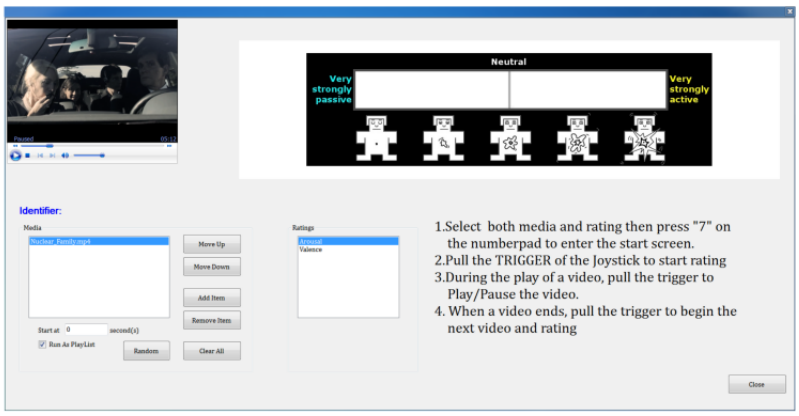
Dans notre protocole expérimental, chaque film peut être visionné par un annotateur seulement une fois car le critère de nouveauté peut influer l'émotion ressentie. Les annotations ont été collectées à partir de 10 participants français (7 femmes et 3 hommes) dont l'age variait entre 18 et 27 ans (moyenne de 21.9 et écart-type de 2.5). Les participants avaient différents niveaux de formation, d'étudiants en Licence à des diplômés de Master. Les expérimentations ont été organisées en 4 sessions, chacune sur une demi-journée. Avant la première session, les participants ont été informés des



(a) Capture d'écran avant l'annotation selon l'axe arousal



(b) Capture d'écran durant l'annotation selon l'axe valence



(c) Menu du programme GTrace modifié

FIGURE 3.4 – Capture d'écran de l'interface d'annotation GTrace modifiée.

objectifs de l'expérience, de la signification des échelles de valence et d'arousal, et ont été entraînés à utiliser l'interface avec trois vidéos de test. Les participants ont annoté les films selon la valence pour les deux premières sessions, et selon l'arousal pour les deux dernières session. Au sein de chaque session, l'ordre des films était

aléatoire.

Finalement, chaque film a été annoté par 5 annotateurs selon la valence, et 5 autres annotateurs selon l'arousal.

Définir des annotations fiables à partir des auto-évaluations continues des différents annotateurs est un point crucial pour permettre l'apprentissage et l'évaluation de modèles performants de prédiction de l'émotion. Deux aspects sont particulièrement importants : chaque annotateur est caractérisé par un certain délai d'annotation (temps entre l'émotion ressentie et son indication avec le joystick), et l'agrégation des auto-évaluations de multiples annotateurs doit prendre en compte la variabilité des annotations [MN13]. Ainsi, plusieurs techniques ont été proposées et étudiées dans la littérature pour gérer la synchronisation des évaluations de plusieurs individus. Dans notre travail, nous avons combiné et adapté les approches proposées par Mariooryad et Busso [MB13] et par Nicolaou et al. [NGP11] pour gérer à la fois les délais d'annotation et la variabilité.

Premièrement, les auto-évaluations enregistrées à une fréquence de 100 valeurs par secondes sont sous-échantillonnées en moyennant les annotations sur des fenêtres de 10 secondes avec un déplacement d'une seconde (donc 1 valeur par seconde). Ce processus permet d'éliminer une grande partie du bruit dû à des mouvements involontaires du joystick. De plus, en raison de la granularité des émotions, une valeur par seconde est suffisante pour représenter les émotions suscitées par les films [MN13].

Ensuite, chaque auto-évaluation est translatée de manière à ce que les annotations translatées de τ maximisent l'accord inter-annotateurs entre l'auto-évaluation translatée de τ et les auto-évaluations non translatées des autres annotateurs. L'accord inter-annotateurs est mesuré par le «Randolph's kappa» [Ran10]. En pratique, τ varie entre 0 et 6 secondes, mais les valeurs les plus élevées (5 à 6 secondes) sont rarement observées (moyenne de 1,47 secondes et écart-type de 1,53 secondes). Finalement, afin d'agréger les différentes annotations, nous avons utilisé une approche similaire à celle proposée dans [NGP11]. La corrélation inter-codeurs est utilisée afin d'obtenir une mesure indiquant la similarité entre l'auto-évaluation d'un codeur et les auto-évaluations des autres participants. La corrélation inter-codeur est définie comme la moyenne des coefficients de corrélation des rangs de Spearman (SRCC) entre les annotations d'un codeur et chacune des annotations des autres codeurs. Cette corrélation inter-codeurs peut alors être utilisée comme pondération lors de la combinaison des annotations de plusieurs annotateurs.

Les Figures 3.5 et 3.6 montrent respectivement les annotations brutes et post-traitées pour l'arousal et la valence pour le film «Spaceman». Les courbes en gras sont

les moyennes pondérées des annotations continues calculées à partir des annotations brutes ou celles lissées et translatées.

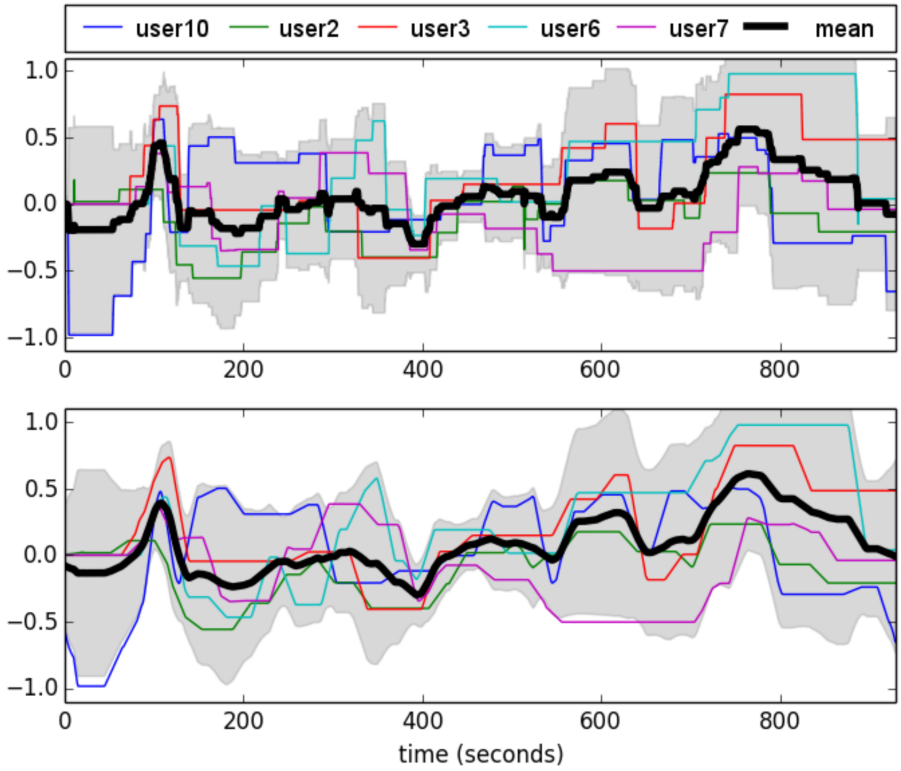


FIGURE 3.5 – Annotations brutes et post-traitées pour l'arousal du film «Spaceman». La zone grise représente l'intervalle de confiance de la moyenne à 95%. Les courbes de la partie supérieure correspondent aux annotations brutes, et celles de la partie inférieure, aux annotations post-traitées.

Pour conclure, ce post-traitement assigne une valeur de valence attendue et une valeur d'arousal attendue pour chaque segment d'une seconde d'un film. Ces valeurs sont normalisées pour varier entre 0 et 1.

Collection «MediaEval 2015 Affective Impact of Movies»

Cette collection a été utilisée comme ensemble de développement et de test pour la tâche «Affective Impact of Movies» à MediaEval 2015 [SBW+15]. Le cas d'utilisation visé était le développement d'un système de recherche vidéo utilisant des outils automatiques pour aider les utilisateurs à trouver des vidéos adaptées à leur humeur, leur age ou préférences. Pour traiter ce problème, deux sous-tâches ont été proposées :

— Détection de l'émotion attendue : l'impact émotionnel d'une vidéo ou d'un film est un indicateur important pour la recherche ou la recommandation ;

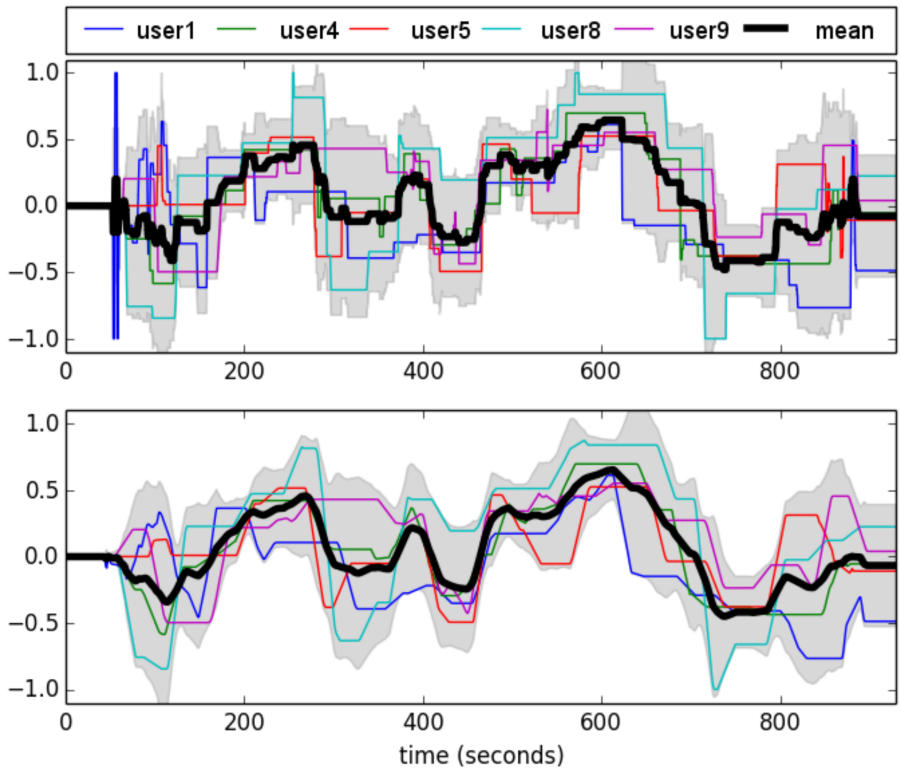


FIGURE 3.6 – Annotations brutes et post-traitées pour la valence du film «Spaceman». La zone grise représente l'intervalle de confiance de la moyenne à 95%. Les courbes de la partie supérieure correspondent aux annotations brutes, et celles de la partie inférieure, aux annotations post-traitées.

— Détection de la violence : détecter un contenu violent est un aspect important pour le filtrage de contenus vidéo basé sur l'age.

Les 9800 clips vidéo de la collection «Discrete LIRIS-ACCEDE» ont été utilisés comme ensemble de développement, et un nouvel ensemble de 1100 clips vidéo a été constitué et utilisé comme ensemble de test. Ainsi, pour chacun des 10900 clips vidéo, l'annotation consiste en une valeur binaire indiquant la présence de violence, la classe du clip pour l'arousal attendue (calme-neutre-active) et la classe pour la valence attendue (négative-neutre-positive).

L'annotation de violence a été réalisée de la manière suivante. Tout d'abord, toutes les vidéos ont été annotées séparément par deux groupes d'annotateurs de deux pays différents. Pour chaque groupe, des annotateurs standards (étudiants diplômés, typiquement célibataires et sans enfants) ont annoté toutes les vidéos, qui ont ensuite été vérifiées par des annotateurs seniors (chercheurs confirmés). Dans chaque groupe, chaque vidéo a reçu deux annotations qui ont ensuite été fusionnées par les annotateurs seniors pour obtenir l'annotation finale pour le groupe. Finalement, les annotations des deux groupes ont été fusionnées et vérifiées une nouvelle

fois par les organisateurs de la tâche.

Afin de rendre les deux sous-tâches compatibles et permettre aux participants d'utiliser des systèmes similaires dans les deux cas, les valeurs de valence et d'arousal ont été discrétisées en trois classes (calme-neutre-active) pour l'arousal et (négative-neutre-positive) pour la valence. Les scores de valence et d'arousal normalisés variant entre -1 et +1, les classes négative et calme correspondent respectivement aux clips vidéo avec un score de valence ou d'arousal inférieur à -0,15, la classe neutre pour les deux dimensions correspond à des scores entre -0,15 et +0,15 et les classes positives et actives correspondent à des scores supérieurs à +0,15. Ces limites ont été fixées empiriquement en prenant en compte la distribution des données dans l'espace valence-arousal.

Les détails pour ces annotations peuvent être trouvés dans [SBW+15].

Collection «MediaEval 2016 Emotional Impact of Movies»

Cette collection a été utilisée comme ensemble de développement et de test pour la tâche «Emotional Impact of Movies» à MediaEval 2016 [DCB+16]. Celle-ci propose aux participants de développer des systèmes afin de prédire automatiquement l'impact émotionnel de films en termes de valence et d'arousal.

Deux sous-tâches ont été proposées :

— Prédiction globale de l'émotion : étant donné un court clip vidéo (environ 10 secondes), les systèmes des participants sont supposés prédire un score de valence et d'arousal attendues pour l'ensemble du clip ;

— Prédiction continue de l'émotion : dans la mesure où une émotion ressentie durant une scène peut être influencée par les émotions ressenties précédemment, l'objectif ici est de considérer des vidéos plus longues et de prédire la valence et l'arousal de manière continue tout au long de la vidéo. Ainsi, un score de valence et d'arousal attendues doivent être fournis pour chaque segment d'une seconde des vidéos.

L'ensemble de développement est composé de la collection «Discrete LIRIS-ACCEDE» pour la première sous-tâche, et de la collection «Continuous LIRIS-ACCEDE» pour la deuxième sous-tâche. Pour constituer les ensembles de test 49 nouveaux films ont été sélectionnés. En suivant le même protocole d'annotation que pour les collections «Discrete LIRIS-ACCEDE» et «Continuous LIRIS-ACCEDE», 1200 nouveaux clips vidéos ont été extraits et annotés pour la première sous-tâche (entre 8 et 12 secondes) alors que 10 films longs (de 25 minutes à 1 heure et 35 minutes) ont été sélectionnés et annotés pour la deuxième sous-tâche (pour une durée

totale de 11,48 heures).

Les détails pour ces annotations peuvent être trouvés dans [DCB+16].

Collection «MediaEval 2017 Emotional Impact of Movies»

Cette collection a été utilisée comme ensemble de développement et de test pour la tâche «Emotional Impact of Movies» à MediaEval 2017 [DHC+17]. Ici, seuls des films longs ont été retenus et l'émotion a été considérée en termes de valence, arousal et peur. Les sous-tâches suivantes ont été proposées, pour lesquelles l'impact émotionnel devait être prédit pour des segments consécutifs de 10 secondes couvrant l'ensemble du film avec une translation de 5 secondes :

— Prédiction de la valence et de l'arousal : les systèmes des participants sont supposés prédire un score de valence et d'arousal attendues pour tous les segments consécutifs de 10 secondes ;

— Prédiction de la peur : l'objectif ici est de prédire si les segments de 10 secondes sont susceptibles de susciter ou non la peur. Le cas d'utilisation visé est la prédiction de scènes effrayantes pour aider le systèmes à protéger les enfants de contenus vidéos non appropriés.

La collection «Continuous LIRIS-ACCEDE» a été utilisée comme ensemble de développement pour les deux sous-tâches. L'ensemble de test consiste en une sélection de 14 nouveaux films sous licence Creative Commons, en supplément des 160 films originaux. Leur durée varie entre 210 et 6260 secondes, et la durée totale est de 7 heures, 57 minutes et 13 secondes. En supplément des données vidéos, des descripteurs visuels et sonores ont également été fournis. Les annotations consistent donc en un score de valence et d'arousal pour chaque segment de 10 secondes (en suivant le même protocole que pour la collection «Continuous LIRIS-ACCEDE»), ainsi qu'une valeur binaire indiquant si le segment est supposé susciter ou non la peur.

L'annotation de la peur a été générée en utilisant un outil spécifique conçu pour la classification de media audio-visuels, permettant de réaliser l'annotation pendant le visionnage du film. Les annotations ont été réalisées par deux membres expérimentés de NICAM, entraînés à la classification de médias. Chaque film a été annoté par un annotateur indiquant les instants de début et de fin des séquences dans le film susceptibles de susciter la peur.

Les détails pour ces annotations peuvent être trouvés dans [DHC+17].

Collection «MediaEval 2018 Emotional Impact of Movies»

Cette collection a été utilisée comme ensemble de développement et de test pour la tâche «Emotional Impact of Movies» à MediaEval 2018 [DHC+18]. Cette tâche était semblable à celle de 2017. Cependant, dans ce cas, plus de données ont été fournies et l'impact émotionnel devait être prédit pour chaque seconde du film, plutôt que pour des segments consécutifs de 10 secondes comme précédemment. Les deux sous-tâches étaient les suivantes :

— Prédiction de la valence et de l'arousal : les systèmes des participants sont supposés prédire un score de valence et d'arousal de manière continue (toutes les secondes) pour chaque film ;

— Prédiction de la peur : l'objectif ici est de prédire l'instant de début et de fin des séquences susceptibles de susciter la peur dans le film. Tout comme l'année précédente, le cas d'utilisation visé était la prédiction de scènes effrayantes pour aider le systèmes à protéger les enfants de contenus vidéos non appropriés.

L'ensemble de développement est constitué des données provenant de la collection «Continuous LIRIS-ACCEDE» ainsi que celles provenant de l'ensemble de test de la collection «MediaEval 2017 Emotional Impact of Movies», soit un total de 44 films pour une durée totale de 15 heures et 20 minutes. L'ensemble de test consiste en 12 films sélectionnés à partir des 160 films d'origine et non encore utilisés pour une annotation continue, pour une durée totale de 8 heures et 56 minutes. Comme pour la collection de 2017, des descripteurs visuels et sonores sont également fournis.

Les annotations ont été obtenues en suivant le même protocole que pour les collections précédentes, et consistent donc en des scores de valence et d'arousal pour chaque seconde des films (pour la première sous-tâche), ainsi que les instants de début et de fin des séquences dans les films susceptibles de susciter la peur (pour la deuxième sous-tâche).

Les détails pour ces annotations peuvent être trouvés dans [DHC+18].

3.3 Modèle spatio-temporel profond pour la prédiction de l'impact émotionnel des vidéos

En s'appuyant sur la plateforme de données LIRIS-ACCEDE présentée précédemment, nous avons proposé plusieurs modèles pour la prédiction des émotions suscitées par les vidéos. En particulier, l'un deux est un modèle spatio-temporel profond visant à intégrer des propriétés psychologiques des émotions.

En effet, les psychologistes suggèrent que l'évaluation d'une émotion est un processus itératif. Par exemple, Russel déclare pour définir le «Core Affect» que [Rus03] :

«Emotional life consists of the continuous fluctuations in core affect, in pervasive perception of affective qualities, and in the frequent attribution of core affect to a single Object, all interacting with perceptual, cognitive, and behavior processes.»

Ce processus d'évaluations récursives et continues est également au coeur de l'«appraisal evaluation» postulée par Scherer [SK10] :

«In the case of humans, the CPM postulates that the recursive checking process repeats the sequence continuously, constantly updating the appraisal results that change rapidly with changing events and evolving [emotional] evaluation until the monitoring subsystem signals termination of or adjustment to the stimulation that originally elicited the appraisal episode.»

Ainsi, cet aspect fondamental des émotions devrait être pris en compte pour l'élaboration de modèles de prédiction de l'émotion. Nous avons donc proposé un modèle intégrant ces notions pyschologiques en nous appuyant sur la dimension temporelles des données vidéo. En d'autres termes, les émotions suscitées précédemment doivent être prises en compte et intégrées récursivement dans le modèle. Ainsi, nous avons développé un modèle profond statique permettant d'extraire des représentation d'images de niveau intermédiaire afin d'alimenter un réseau de neurones récurrent de type «Long Short-Term Memory» [HS97] capable de modéliser les relation émotionnelles entre des segments consécutifs de vidéos.

Nous avons considéré dans ce travail le modèle profond GoogleNet introduit dans [SWY+15] devenu en 2014 la nouvelle référence avec les meilleures performances sur le jeu de données ImageNet. Il consiste en la concaténation de réseaux «Inception» qui sont eux-mêmes constitués de convolutions de taille 1x1, 3x3, 5x5 empilés avec des couches de max-pooling pour réduire la résolution. Compte-tenu de la profondeur importante du réseau, deux fonctions de perte auxiliaires sont connectées à des couches intermédiaires pour augmenter le gradient retro-propagé et pour éviter le problème de la disparition du gradient. Durant l'apprentissage, les fonctions de perte auxiliaires pondérées sont ajoutées à la perte totale du réseau. Pour le test par contre, ces fonctions ne sont plus nécessaires et sont donc supprimées du réseau. Dans nos expériences, les activations softmax auxiliaires et finale sont remplacées par une couche entièrement connectée composée d'un unique neurone avec une activation sigmoïde pour produire la prédiction du score de valence ou d'arousal.

Le réseau pré-entrainé sur ImageNet est ensuite ajusté sur nos données d'apprentissage de la collection «Continuous LIRIS-ACCEDE» présentée dans la section 3.2. La dernière couche entièrement connectée fournit alors une représentation compacte

(1024 neurones) de niveau intermédiaire des données adaptée à la prédiction des émotions, et qui pourra ensuite être utilisée par le modèle temporel.

Afin de prendre en compte la modalité sonore des vidéos qui joue une part importante dans l'impact émotionnel des vidéos, nous avons utilisé les spectrogrammes audio qui peuvent être considérés comme des représentations visuelles du contenu sonore et donc alimenter des réseaux convolutifs. Cette stratégie s'est révélée être efficace notamment pour la reconnaissance de l'émotion de parole ou encore la détection d'onset dans la musique [HDMZ14, SB14]. Ainsi, les signaux sonores des canaux gauche et droit extraits à partir de segments vidéo d'une durée d'1 seconde sont tout d'abord convertis en spectrogrammes par application de la transformée de Fourier à court terme [All77]. Les deux spectrogrammes sont ensuite additionnés et l'image obtenue est redimensionnée pour obtenir une taille de 256x256 pixels, utilisée comme donnée d'entrée d'un réseau convolutif basé audio. Trois exemples de tels spectrogrammes sont présentés dans la Figure 3.7.

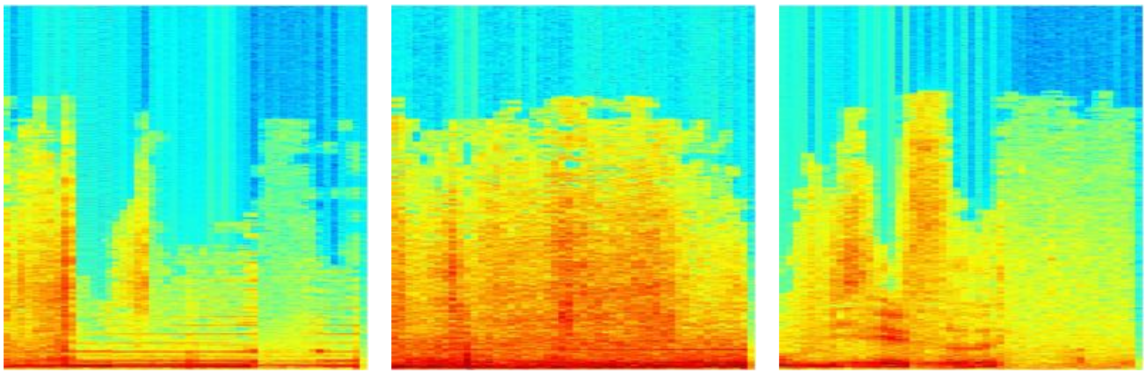


FIGURE 3.7 – Exemples de spectrogrammes redimensionnés utilisés comme entrée de CNN basés audio.

Différents modèles ont été proposés dans la littérature pour prendre en compte la temporalité, en particulier les réseaux de neurones récurrents (RNN) [RHW86]. Afin de pouvoir capturer des dépendances à long terme, ils peuvent être combinés avec des unités Long Short-Term Memory (LSTM) [HS97]. Le principe de ces LSTM-RNNs est ainsi de remplacer les unités cachés usuelles des réseaux de neurones par ces unités LSTM pouvant apprendre quand mémoriser et quand oublier les dépendances passées. Ces réseaux ont été utilisés dans plusieurs travaux liés aux émotions, notamment pour la prédiction de la valence et de l'arousal à partir de données audiovisuelles et physiologiques [REK+15] ou encore pour la regression continue de l'émotion dans la musique [WES14]. Ces applications fructueuses nous ont donc conforté dans le choix d'étudier la capacité des réseaux récurrents LSTM-RNNs à modéliser l'aspect temporel des émotions suscitées par les vidéos.

La Figure 3.8 illustre une unité LSTM.

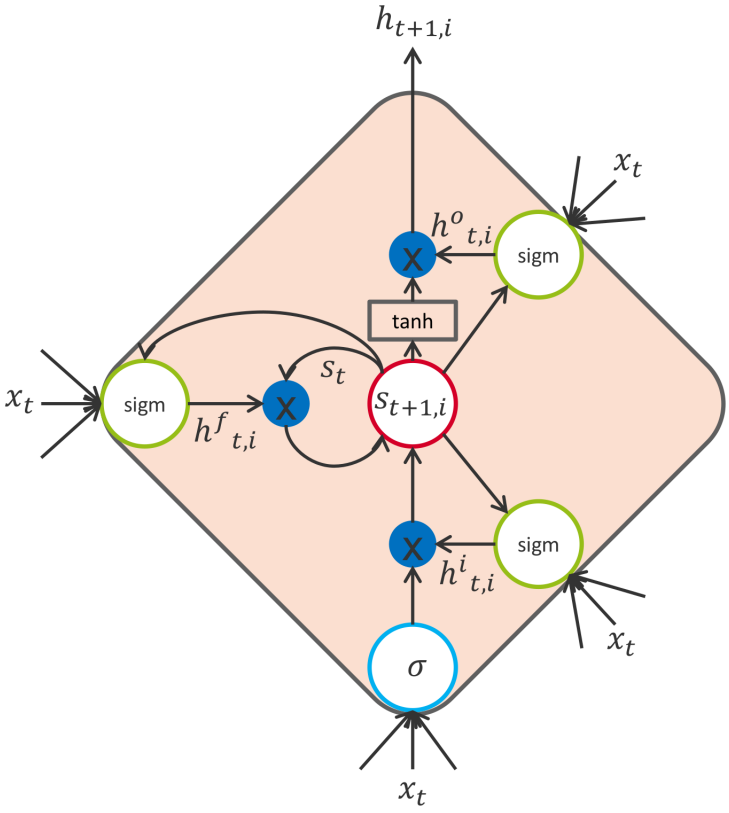


FIGURE 3.8 – Schéma d'une unité LSTM. Le cercle rouge représente la cellule d'état et les trois cercles verts représentent les portes d'entrée, d'oubli et de sortie.

La mémoire d'une unité LSTM i pour l'instant t est gérée par trois portes sigmoïdes : la porte d'entrée $h^i_{t,i}$, la porte d'oubli $h^f_{t,i}$ et la porte de sortie $h^o_{t,i}$. Les portes sont activées en transmettant leur entrée à travers une fonction sigmoïde. Le composant le plus important d'une unité LSTM est la cellule d'état s_t qui a une boucle linéaire sur elle-même pondérée par la porte d'oubli. La sortie de la cellule d'état est transmise à travers une fonction *tanh* et peut être éteinte avec la porte de sortie. Ces valeurs sont calculées de la manière suivante :

$$h^f_{t,i} = sigmoid(b^f_i + \sum_j U^f_{ij} x_{t,j} + \sum_j V^f_{ij} s_{t,j} + \sum_j Wij^f h_{t,j}) \tag{3.1}$$

$$s_{t+1,i} = h^f_{t,i} s_{t,i} + h^i_{t,i} \sigma(b_i + \sum_j U_{ij} x_{t,j} + \sum_j Wij h_{t,j}) \tag{3.2}$$

$$h^i_{t,i} = sigmoid(b^i_i + \sum_j U^i_{ij} x_{t,j} + \sum_j V^i_{ij} s_{t,j} + \sum_j Wij^i h_{t,j}) \tag{3.3}$$

$$h^o_{t,i} = sigmoid(b^o_i + \sum_j U^o_{ij} x_{t,j} + \sum_j V^o_{ij} s_{t,j} + \sum_j Wij^o h_{t,j}) \tag{3.4}$$

$$h_{t+1,i} = tanh(s_{t+1,i}) h^o_{t,i} \tag{3.5}$$

où x_t est le vecteur d'entrée courant, h_t est le vecteur de la couche cachée courante, σ est une fonction d'activation (sigmoïde ou tanh par exemple), b^f, b^i, b^o sont les biais pour les portes, U^f, U^i, U^o sont les poids d'entrée pour les portes et V^f, V^i, V^o, W^f, W^i, W^o sont les poids récurrents pour les trois portes.

Afin d'intégrer la temporalité dans notre modèle, nous nous sommes appuyés sur le principe des LSTM-RNNs bidirectionnels [GS05] qui sont la combinaison des concepts des LSTM-RNNs présentés précédemment et des RNN bidirectionnels [SP97]. L'avantage d'un RNN bidirectionnel est qu'il a non seulement accès au informations passées, mais également aux informations futures grâce à l'utilisation de deux couches d'entrée distinctes traitant les données en avant et en arrière. Un LSTM-RNN bidirectionnel a donc la capacité d'accéder aux dépendances à long terme dans les deux directions.

Le principe général de notre modèle spatio-temporel de prédiction de l'émotion est illustré dans la Figure 3.9. Les étapes permettant de créer une courbe émotionnelle (série de valeurs de valence ou d'arousal) d'une vidéo sont les suivantes :

1. La vidéo est tout d'abord segmentée en segments vidéo consécutifs d'une durée d'1 seconde.
2. Un spectrogramme audio et un patch d'une image clé sont extraits du premier segment vidéo.
3. Le spectrogramme est utilisé pour alimenter un CNN audio entraîné pour l'arousal ou la valence en fonction de la sortie désirée du modèle spatio-temporel.
4. En parallèle le patch est utilisé pour alimenter un CNN visuel entraîné lui aussi pour l'arousal ou la valence
5. Les représentations de niveau intermédiaire sont extraits à partir des deux CNN (activations de la dernière couche entièrement connectée) et leur dimen-

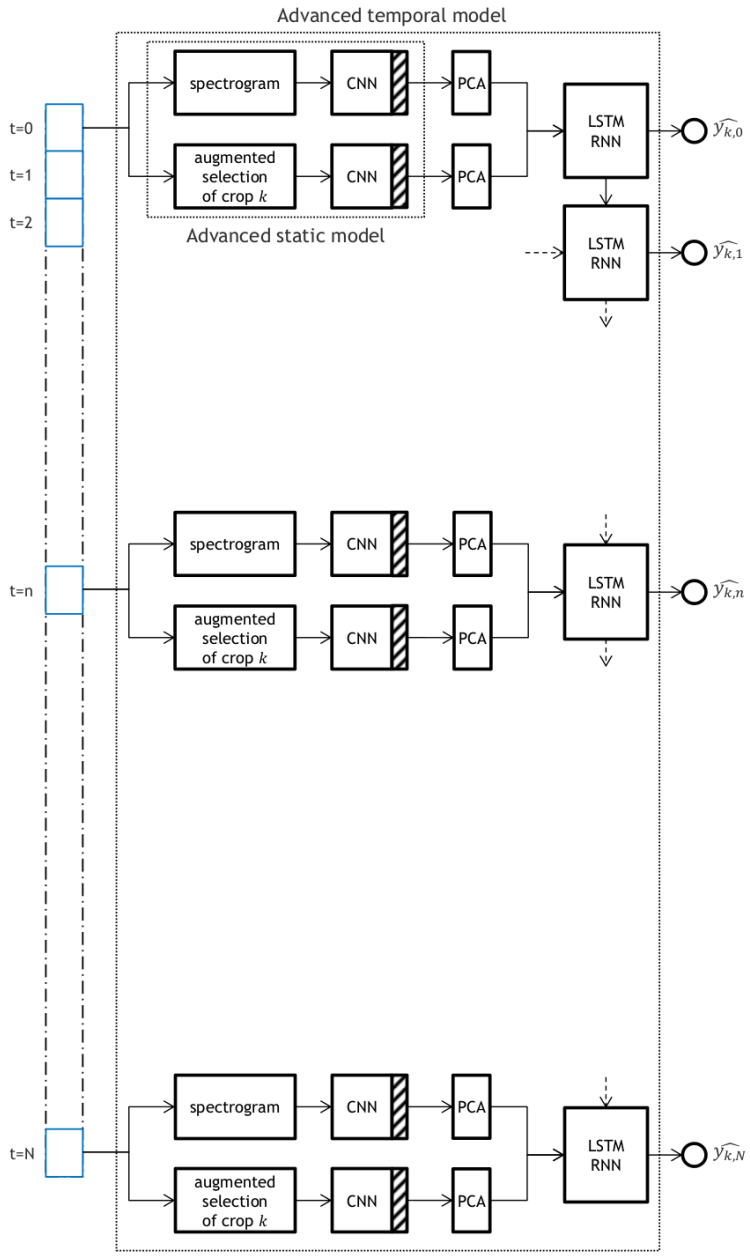


FIGURE 3.9 – Principe général de notre modèle spatio-temporel pour la prédiction d'une courbe émotionnelle d'une vidéo.

sion est réduite par l'application d'une Analyse en Composantes Principales. Les deux modalités sont alors combinées pour former une unique représentation compacte multimodale de niveau intermédiaire à partir du segment vidéo.

6. Cette représentation est utilisée pour alimenter le LSTM-RNN bidirectionnel pour produire en sortie un score émotionnel (valence ou arousal).

7. Les étapes 2 à 6 sont répétées pour le segment vidéo d'1 seconde suivant, et ce, jusqu'à ce que le dernier segment ait été traité.

8. Finalement, la courbe émotionnelle obtenue, représentant l'évolution de la valence ou de l'arousal au cours de la vidéo, est lissée par application d'un filtre Gaussien.

Comme l'information visuelle est portée par un patch extrait d'une image clé d'un segment d'une seconde, afin d'augmenter la robustesse du modèle, lors de la phase de prédiction, cinq courbes émotionnelles sont calculés de manière à ce que les patchs sélectionnés soit différents. Afin de combiner ces courbes, deux stratégies ont été étudiées. La première est une simple moyenne alors que la seconde est une moyenne pondérée par des poids calculés pour maximiser les performances sur un ensemble de validation.

Les détails de l'apprentissage des CNN audio et visuel, ainsi que du LSTM-RNN peuvent être trouvés dans [Bav15].

Les expérimentations avec notre modèle spatio-temporel pour la prédiction de l'émotion ont été conduites sur la collection «Continuous LIRIS-ACCEDE» présentée dans la section précédente. Sur les 30 vidéos annotées disponibles, 27 ont été utilisées comme ensemble de développement, et 7 comme ensemble de test. Les performances du modèle spatio-temporel ont été comparées à celles de modèles statiques (n'intégrant pas la temporalité). Comme il s'agit d'un problème de régression (production d'un score continu de valence et d'arousal), les indicateurs de performance retenus sont l'erreur quadratique moyenne (MSE) traditionnellement utilisée dans ces situations, et également le coefficient r de Pearson mesurant la corrélation linéaire entre la série de valeurs prédite et la série des valeurs réelles.

Les résultats sont donnés dans la Figure 3.10. Le modèle «SVR with transfer learning baseline» correspond à un SVR (Support Vector Regressor) alimenté par des descripteurs CNN (couche $fc7$ d'un réseau AlexNet [KSH12]) pré-entraîné sur ImageNet et ajusté sur notre jeu de données LIRIS-ACCEDE. Le modèle «Multimodal static» correspond à la combinaison des CNN audio et visuel mentionnés précédemment, directement utilisés pour la prédiction et non comme entrée d'un LSTM-RNN. Avec une corrélation de 0,559 obtenue pour la valence et de 0,361 pour l'arousal, notre modèle spatio-temporel se révèle être plus performant que les modèles statiques, et bien supérieur à un modèle aléatoire ou uniforme. Cependant, comme on peut le voir, la marge d'amélioration est encore importante, particulièrement pour l'arousal. Ainsi, même si ce modèle spatio-temporel est une étape importante dans la modélisation des émotions, le chemin à parcourir est encore long pour disposer de modèles très performants.

Une analyse plus approndie des résultats est disponible dans [Bav15].

Model	Arousal		Valence	
	MSE	r	MSE	r
Random	0.109	0.0004	0.113	-0.002
Uniform	0.026	-0.016	0.029	-0.005
SVR with transfer learning baseline	0.022	0.337	0.034	0.296
Multimodal static model (best fusions)	0.018	0.170	0.027	0.349
Multimodal static model with Gaussian smoothing	0.020	0.208	0.025	0.489
LSTM-RNN (simple average)	0.018	0.289	**0.024**	**0.559**
LSTM-RNN (best weights)	**0.017**	**0.361**	**0.024**	**0.559**

FIGURE 3.10 – Performance de notre modèle spatio-temporel pour la prédiction de l'émotion comparée à celle de modèles statiques (n'intégrant pas de temporalité).

3.4 Conclusion

Ce chapitre retrace nos principales contributions pour la prédiction de l'impact émotionnel des vidéos.

Tout d'abord, afin de répondre aux besoins de la communauté qui ne disposait pas de jeu de données suffisamment riche et varié pour réaliser des expérimentations avancées sur les émotions suscitées par les vidéos, nous avons proposé la plateforme de données LIRIS-ACCEDE constituée de six collections : les deux collections historiques «Discrete LIRIS-ACCEDE» et «Continuous LIRIS-ACCEDE» ainsi que quatre collections régulièrement ajoutées pour enrichir la plateforme, en particulier dans le cadre de son utilisation pour la tâche «Emotional Impact of Movies» à MediaEval. Son intérêt réside dans le fait que les vidéos utilisées sont sous licence «Creative Commons», ce qui permet de les distribuer librement, et que le nombre de vidéos est très important : sur l'ensemble des collections, 10900 extraits vidéos d'une dizaine de secondes sont annotés globalement selon la valence et l'arousal, et 66 films (36 heures) sont annotés de manière continue (chaque seconde) selon la valence, l'arousal, mais également la peur.

De plus, en utilisant cette plateforme de données, nous avons pu développer et évaluer plusieurs modèles de prédiction de l'émotion suscitée par les vidéos. En particulier, nous avons proposé une modèle spatio-temporel afin d'intégrer la temporalité des émotions. Ce modèle repose sur deux réseaux convolutifs dédiés à la modalité visuelle et à la modalité sonore. Les informations multimodales fournies

par ces deux réseaux sont ensuite utilisée par un réseau LSTM-RNN bidirectionnel pour modéliser les dépendances temporelles entre les segments vidéo successifs des films. Les résultats expérimentaux ont montré un bonne capacité de ce modèle à prédire les émotions, même si la marge de progression est encore importante pour avoir une prédiction suffisamment fiable pour être utilisée par des systèmes commerciaux visant par exemple la recommandation de vidéos basées sur l'émotion, ou encore le filtrage de contenu non adaptés à certains publics.

Chapitre 4

Analyse visuelle pour la robotique

4.1 Introduction

Une deuxième application privilégiée de nos travaux concerne la robotique. En effet, la robotisation croissante de tâches pénibles et répétitives est un symbole de progrès technologique au service de l'homme. Ainsi, pour permettre une robotisation croissante de tâches de plus en plus complexes, mais souvent fastidieuses et/ou dangereuses pour les êtres humains, il est nécessaire de doter les robots d'une vision artificielle qui leur permette d'observer et de comprendre la scène, ainsi que d'une intelligence leur permettant d'acquérir de nouvelles capacités ou de s'adapter aux changements d'environnements. Dans ce cadre, nous développons en particulier des méthodes d'apprentissage automatique et de vision par ordinateurs pour créer des outils de Picking/Kitting (prélèvement et dépose d'objets) sur des bases robotiques afin de les rendre flexibles, adaptables et autonomes.

Ces travaux ont été réalisés avec les doctorants Matthieu Grard et Amaury Depierre, notamment dans le cadre du projet FUI Pikaflex et du Labcom Arès, en étroite collaboration avec l'entreprise Siléane. Ils ont été publiés dans [C1, C3] (numérotation correspondant aux publications listées dans le rapport d'activité en fin de document) et font l'objet de deux soumissions : à la revue IJCV (2^{nde} révision), et à la conférence ICRA 2020.

4.2 Localisation d'instances d'objets dans un vrac

La délimitation d'instances d'objets et la compréhension de leur disposition spatiales à partir d'une unique image RGB sans modèle explicite des objets est une tâche de vision par ordinateur au cœur de nombreuses applications telle que la conduite autonome dans des environnements inconnus, ou encore, pour ce qui nous intéresse plus particulièrement ici, pour le picking (la saisie) d'objets par des bras robotiques. En effet, les instances d'objets les moins occultées sont généralement celles à privilégier pour la saisie. Automatiser une telle tâche demeure très compliqué dans la

mesure où le robot doit gérer de nombreuses variations de scènes à partir d'une simple grille de valeurs de couleur RGB.

Les réseaux profonds entièrement convolutionnels (FCN) sont devenus la référence pour l'apprentissage de représentations d'images généralisables, en particulier en raison de leur capacité à capturer des invariants multi-échelle. Dans ce contexte, la tendance pour détecter des instances saillantes consiste à segmenter l'image en la découpant en de nombreuses régions. Un FCN est entraîné à d'abord isoler chaque instance dans une boîte englobante en réalisant simultanément une classification et une regression de boîtes d'ancrage, puis pour chaque boîte, allumer les pixels appartenant à des parties d'instances visibles et non-occultées [ZTMD17, FKH+19, QJL+19] ou à des catégories prédéfinies accessibles [DNR18].

Cependant, approximer une instance avec un rectangle n'est pas toujours pertinent. Typiquement, dans des agencements homogènes denses, tels que des vracs d'objets, de nombreuses instances du même objet s'occultent les unes les autres. Ainsi, une proposition de boîte contient souvent plusieurs instances. Cela est illustré dans la Figure 4.2.

Dans de tels agencements, obtenir une segmentation d'une image ou d'une région conservant les instances devient une tâche difficile puisque une attention au niveau des pixels des instances nécessitent des représentations dépendantes de la position alors que les noyaux convolutifs des réseaux sont invariants à la translation. Généralement, les labels au niveau des pixels sont inférés en combinant graduellement des informations au niveau des objets à faible résolution avec des indices locaux à résolution plus élevée en utilisant un réseau encodeur-décodeur résiduel (RED). Dans une telle structure, le décodeur a pour but de suréchantillonner les représentations latentes de l'encodeur. Les réseaux RED se sont révélés être efficaces en particulier pour inférer les contours d'objets [CZP+18]. Cependant, un encodeur profond peut difficilement être décodé pour distinguer des instances similaires superposées, en raison de son invariance à la translation par nature. La plupart des efforts de recherche pour améliorer la délimitation des objets a été mise sur l'encodeur, en utilisant des couches avec des connections denses pour approfondir les blocs de l'encodeur [HLvdMW17], des convolutions dilatées pour élargir le champ de perception au niveau d'encodage de plus faible résolution [CZP+18] ou des convolutions prenant en compte le coordonnées pour associer les représentations latentes aux positions des pixels [NALV18]. Ces structures mènent à des représentations dépendantes de la position à faible résolution des catégories d'objets, plus facile à suréchantillonner. Cependant, dans des agencements homogènes denses, le processus de décodage a une plus grande importance puisque la diversité des objets à encoder est plus ré-

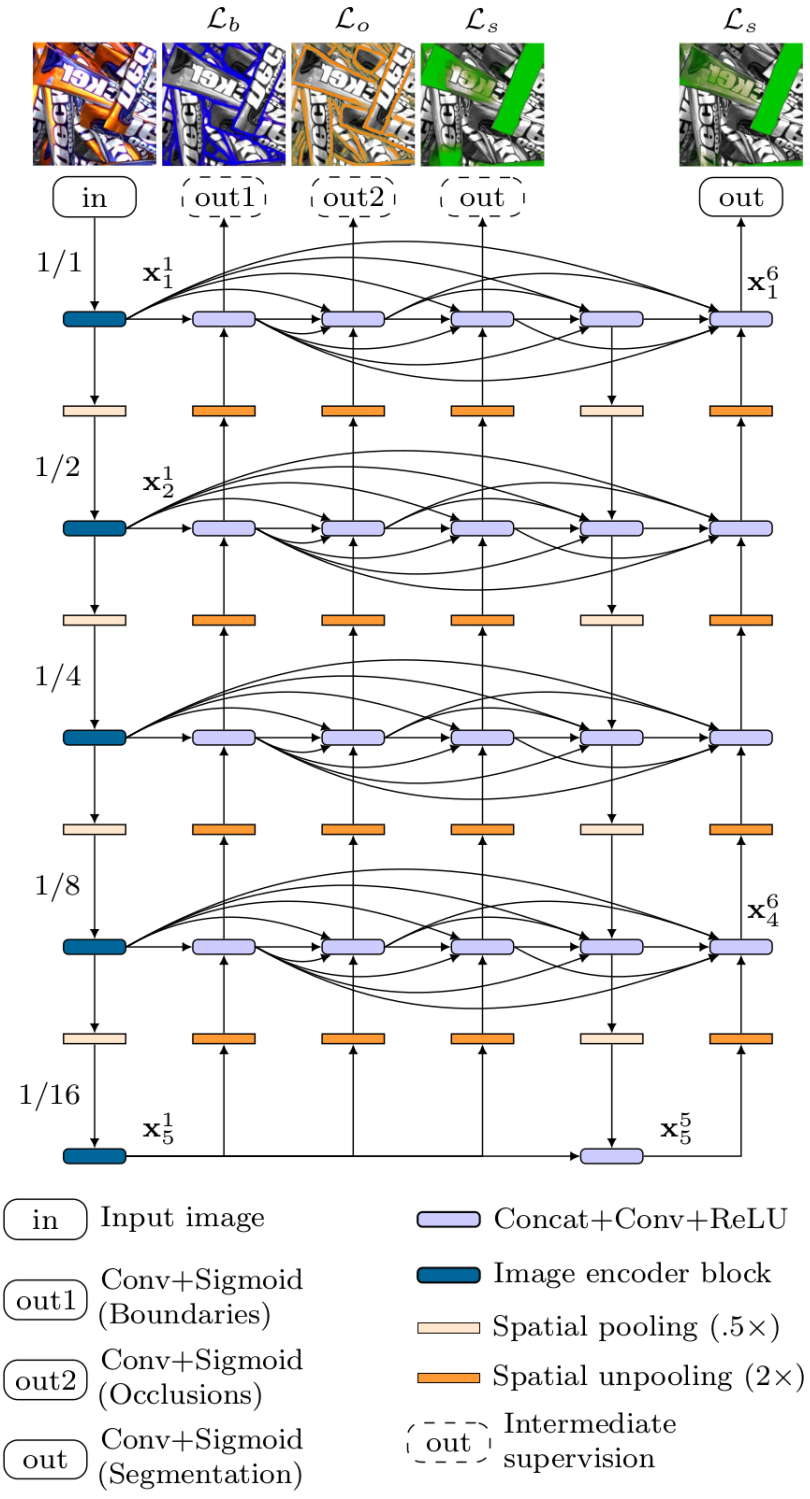


FIGURE 4.1 – Structure de notre modèle (MC6†) pour la segmentation d'instances non occultées.

duite alors que l'encodage des pixels doit permettre de discriminer les instances d'un même objet.

FIGURE 4.2 – Dans des agencements d'objets denses tels que ceux rencontrés dans des images de vracs, les occlusions sont pour la plupart entre instances qui ne peuvent pas être isolées dans un rectangle.

Nous avons donc proposé une nouvelle structure d'encodeur-décodeur résiduel afin de permettre la segmentation d'instances en prenant en compte les occlusions à partir d'une unique image RGB d'agencements homogènes d'instances. Plus précisément, nous avons proposé un processus de décodage plus complexe afin de permettre l'intégration du contexte des pixels pour mieux discriminer les instances similaires.

La structure de notre réseau consiste en un décodeur et des encodeurs-décodeurs légers couplés en cascade de manière dense et supervisés de manière différente pour décomposer la tâche complexe de délimitation des instances non-occultées en tâches plus simples lors du processus de décodage : extraire les informations des images, détecter les contours des instances, détecter le côté occultant du contour des instances (voir la Figure 4.3), allumer les pixels des instances non-occultées et raffiner la segmentation. Ce réseau est illustré dans la Figure 4.1.

Ainsi, à partir d'une image RGB en entrée, le premier élément (colonne de gauche dans la Figure 4.1) est un encodeur profond, par exemple un encodeur VGG16 [SZ15]. Les trois premiers décodeurs en cascade (trois colonnes suivantes) vont graduellement identifier les contours des instances d'objets, le côté intérieur du contour des objets (permet de savoir si l'objet est occulté) et la segmentation mettant en valeur les instances non-occultées. Ces trois unités ont pour objectif de structurer le processus de décodage. Cela permet également la réutilisation d'informations spécifiques aux sous-tâches : l'identification du côté occultant du contour peut s'appuyer

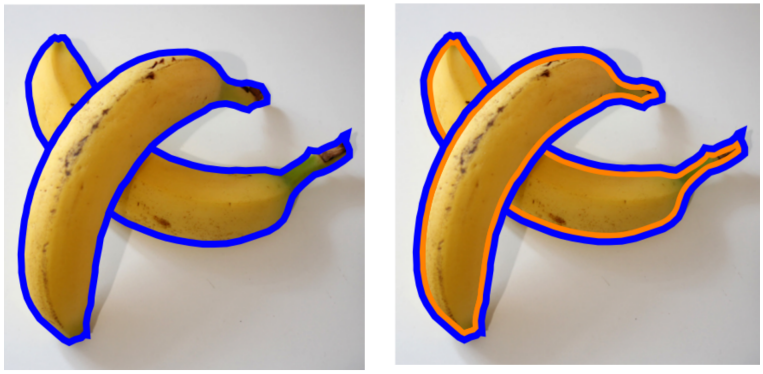


FIGURE 4.3 – Illustration du contour d'instances d'objets et du côté occultant du contour de ces instances. Cette information du côté occultant permet de déduire si un objet est occulté ou non (occulté si le côté occultant est discontinu).

sur l'information de localisation du contour, et la segmentation d'instances non-occultées peut s'appuyer sur le fait que le côté occultant des contours soit continu pour cette instance. A la suite de ces décodeurs, une unité d'encodage-décodage a pour but de raffiner la segmentation. Des connections à saut «skip connections» sont également utilisées entre les couches de l'encodeur et des décodeurs afin de permettre un codage multi-échelle combinant des informations de basse résolution mais de niveau sémantique plus élevé (d'une couche plus profonde) avec des informations de plus haute résolution (d'une couche moins profonde).

Pour entraîner et évaluer notre modèle un large jeu de données d'images est nécessaire. Ces images doivent représenter des agencements d'instances homogènes denses avec de nombreuses occlusions comme cela est rencontré dans de nombreuses applications robotiques de manipulation d'objets. Or, les jeux de données disponibles ne présentaient pas ces caractéristiques de manière satisfaisante (peu d'objets, donc peu d'occlusions). Nous avons donc proposé un protocole afin de réaliser un jeu de données d'images synthétiques annotées, présenté dans la Figure 4.4. Ce jeu de données Mikado, et son extension Mikado+ contiennent au total 16960 images de taille 640x512 avec un nombre d'instances de 507186 soit une moyenne de 30 instances par image.

Les résultats expérimentaux sur le jeu de données Mikado sont donnés dans la Figure 4.5. Trois métriques d'évaluation ont été considérées, typiquement utilisées pour les tâches de classification binaire de pixels (et donc de détection de contours et de segmentation) : ODS, AP et AP_{60}. Ces résultats confirment que la structure du réseau que nous avons développé est plus efficace que les méthodes de référence de l'état de l'art pour capturer des représentations sensibles à la position dans l'image

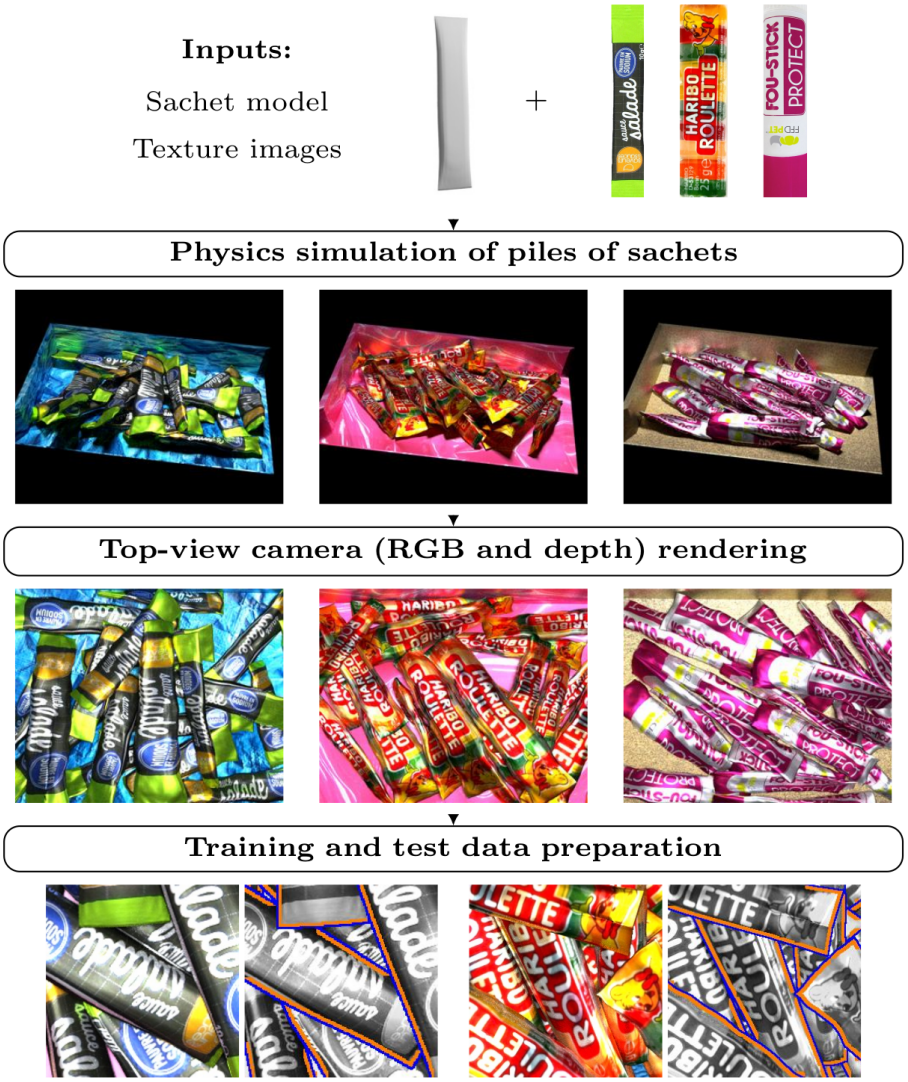


FIGURE 4.4 – Aperçu du procédé utilisé pour la création du jeu de données de vracs intitulé Mikado. A partir d'un maillage et d'images de textures, des vracs d'instances déformées sont générés en utilisant un moteur de simulation physique. Une caméra vue de dessus est ensuite simulée pour obtenir une image RGB et de profondeur. Les images synthétiques et leur annotation (contours en bleu et coté occultant du contour en orange) sont ensuite disponibles pour l'apprentissage et l'évaluation de modèles de détection d'instances non-occultées.

(contours et côté occultant). Plus précisément, notre modèle surpasse RED-Atrous, RED-Coords et RED-Dense/E par 20,6, 7,8 et 5,1 points respectivement en terme d'AP. De plus, comme l'illustre la Figure 4.6, notre réseau fournit des segmentations d'instances de meilleure qualité et des décisions au niveau pixel plus contrastées que les autres modèles.

Des résultats plus détaillés et une analyse approfondie peuvent être trouvés dans

Architecture	Number of parameters	Segmentation ODS	AP	AP_{60}
RED-Atrous	1,957,137	.631	.619	.506
RED-Coords	1,471,105	.703	.747	.599
RED-Dense/E	1,202,217	.724	.774	.593
MC6† (Ours)	5,411,916	**.767**	**.825**	**.691**

FIGURE 4.5 – Résultats comparatifs sur le jeu de données Mikado. Références : RED-Atrous [CZP+18], RED-Coords [LLM+18], RED-Dense/E [HLvdMW17]

[Gra19].

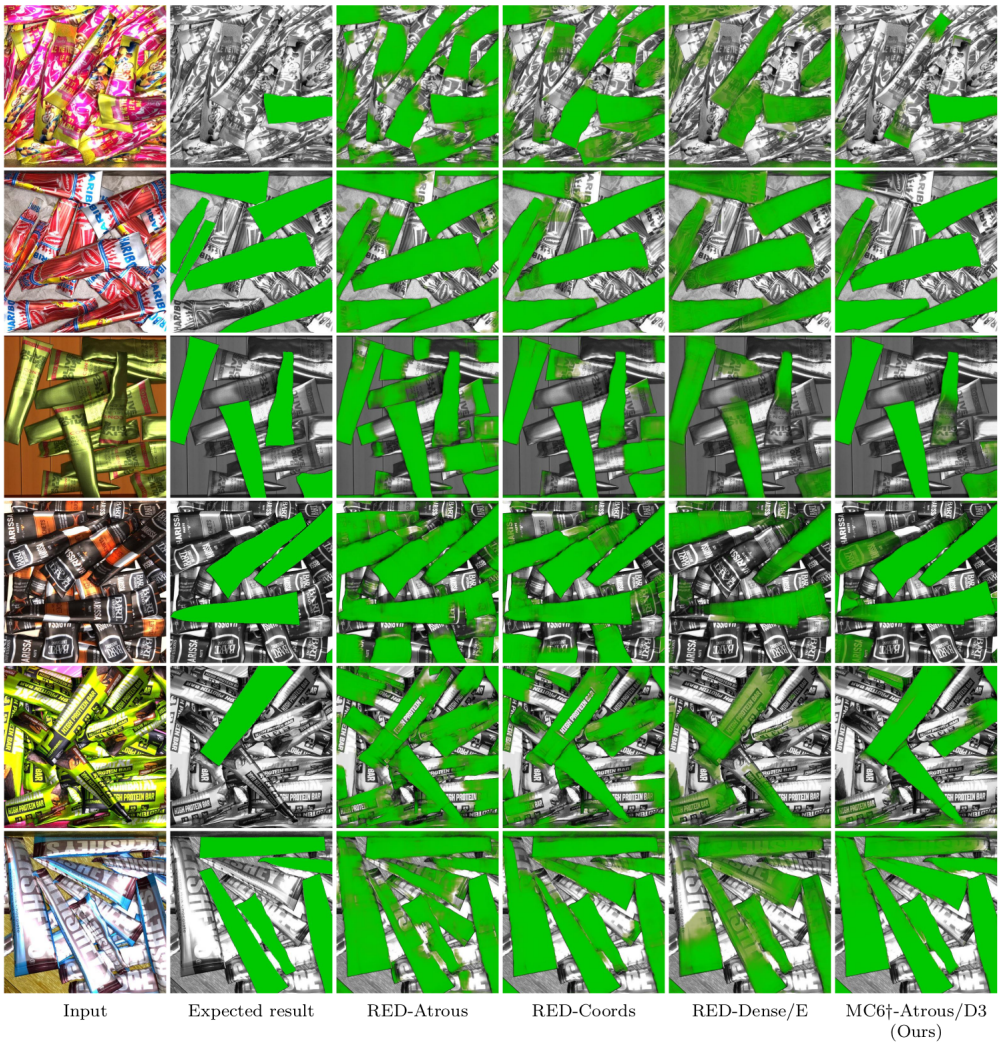


FIGURE 4.6 – Résultats comparatifs sur le jeu de données Mikado. Références : RED-Atrous [CZP+18], RED-Coords [LLM+18], RED-Dense/E [HLvdMW17]

4.3 Prédiction de prises pour des bras robotiques

Au delà de la segmentation des instances les plus prenables, nous nous sommes intéressés justement à la prédiction des paramètres de prises sur les objets présents dans l'image.

En effet, la saisie d'objet est une opération cruciale pour les systèmes robotiques dans de nombreux domaines : industrie et logistique, ménage, interactions avec des humains, ... Un système de saisie robotique efficace et fiable peut mener à une importante amélioration de la productivité, ainsi qu'à de nouvelles applications. Or, les performances obtenues par les systèmes actuels sont encore très loin de celles des humains. Ces derniers peuvent saisir des objets connus et inconnus, de toute forme, dans des conditions d'éclairage sombres ou lumineuses, avec un taux de succès proche de 100%. Pour réaliser la même action finale, un système robotique doit (1) analyser les données fournies par le capteur pour trouver une bonne position de prise, (2) planifier une trajectoire pour atteindre cette position et (3) activer le préhenseur saisir l'objet. Dans nos travaux, nous nous sommes intéressés à la première étape de ce processus et plus précisément à la détection de positions de prises pour une pince à mâchoires parallèles dans des images RGB.

Une des méthodes les plus efficaces à l'heure actuelle est celle proposée par Zhou et al. dans [ZLZ+18]. Elle s'appuie sur des boîtes d'ancrage dans les images avec des orientations multiples. L'angle de prise est alors prédit à partir de cette orientation de référence, tout comme les autres paramètres de prise (position (x, y) et dimension de la pince (w, h)). Le réseau de type entièrement convolutif a alors pour but de prédire par régression 5 valeurs ainsi qu'un score de qualité de prise pour chaque boîte de référence orientée, et ceci pour chaque position dans la carte de descripteurs. La performance obtenue sur le jeu de données Cornell Grasping Dataset [1] est de 97,74%. Cependant, la prédiction de la qualité de la prise dépend uniquement de l'information dans l'image et n'est donc pas directement liée à la prédiction des valeurs de la prise. La méthode que nous proposons étend cette approche en ajoutant une dépendance directe entre la prédiction des valeurs de la prise et l'évaluation du score caractérisant la qualité de la prise.

Dans nos travaux, nous considérons le problème de détection de la position d'une prise à partir d'images RGB d'objets divers disposés sur un plan. Ainsi, nous utilisons une représentation des prises en deux dimensions. Chaque prise peut alors être

1. http ://pr.cs.cornell.edu/grasping/rect_data/data.php

caractérisée par

$$g = \{x, y, w, h, \theta\} \tag{4.1}$$

où (x, y) sont la position du centre de la prise, (h, w) ses dimensions et θ son orientation, comme présenté dans la Figure 4.7(a).

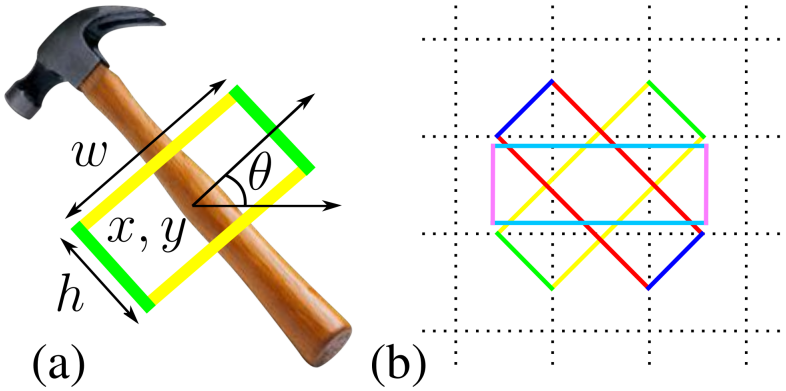


FIGURE 4.7 – (a) Exemple d'une représentation d'une prise avec 5 valeurs. (b) Trois exemples de boîtes d'ancrage orientées de même dimension avec des angles de -45°, 0° et +45° centrées sur le même pixel de la carte de descripteurs.

Afin de simplifier le problème de regression, une connaissance a priori sur la position, la dimension et l'orientation de la prise est introduite par une prise de référence orientée [ZLZ+18]. Ces prises de références, également appelées boîtes d'ancrage sont définies comme $g_a = (g_{ax}, g_{ay}, g_{aw}, g_{ah}, g_{a\theta})$. La prise est alors définie comme une déformation $\delta = (\delta_x, \delta_y, \delta_w, \delta_h, \delta_\theta)$ d'une prise de référence selon les équations suivantes :

$$\begin{aligned} x &= \delta_x * g_{aw} + g_{ax} \\ y &= \delta_y * g_{ah} + g_{ay} \\ w &= exp(\delta_w) * g_{aw} \\ h &= exp(\delta_h) * g_{ah} \\ \theta &= \delta_\theta * (180/k) + g_{a\theta} \end{aligned} \tag{4.2}$$

où k est le nombre de boîtes d'ancrage différentes. La Figure 4.7(b) présente trois exemples différents de boîtes d'ancrage orientées.

Le principe de notre modèle de prédiction de prises est illustré dans la Figure 4.8.

Le premier composant concerne l'extraction de descripteurs. Cela est réalisé pour un réseau ResNet-50 [HZRS16]. Notre choix s'est porté sur cette solution car ce

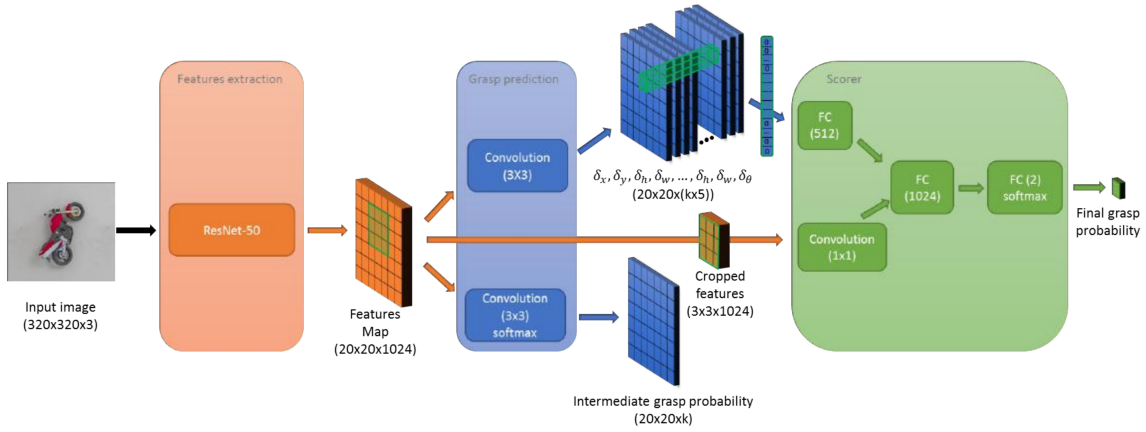


FIGURE 4.8 – Vue globale de l'architecture de notre modèle de prédiction de prises avec ses trois composants : l'extraction de descripteurs, la prédiction de prises et le réseau de score.

réseau s'est montré très efficace dans divers domaines.

Le second composant de notre architecture correspond à un réseau de prédiction de prises basé sur des boîtes d'ancrage orientées, telles que celui proposé par [ZLZ+18]. Deux couches convolutives sont entraînées à prédire pour chaque pixel de la carte de descripteurs et pour chaque boîte d'ancrage de référence un score de classification et cinq valeurs par régression. Le score de classification indique la qualité de la prise de référence orientée correspondante : un score proche de 1 signifie une confiance élevée alors qu'un score proche de 0 indique une position ou une orientation de la pince inadaptée.

Le troisième composant correspond au réseau de score. En effet, le score de qualité de prise intermédiare, fourni par le réseau de prédiction de prises dépend uniquement des descripteurs et de la boîte d'ancrage g_a, et non de la prédiction finale des valeurs de prise g. Ainsi, par exemple, une prise de référence sur une bonne position peut avoir un score élevé malgré une mauvaise prédiction une fois la prise finale g calculée par l'équation 4.2. De plus, l'estimation du score ne peut pas être utilisée pour améliorer la qualité de la régression. Afin d'apporter une solution à ces problèmes, nous avons enrichi le réseau de [ZLZ+18] d'un troisième composant : un réseau de score. En utilisant un sous-ensemble des descripteurs et une représentation de prise, le réseau de score prédit une probabilité décrivant la qualité de la prise proposée.

Trois fonctions de perte sont utilisées dans notre architecture. La fonction de perte de classification pour le réseau de prédiction et pour le réseau de score sont toutes deux des pertes de type softmax cross-entropy. Elles sont calculées unique-

ment sur un sous-ensemble des prédictions : P positives et $3P$ négatives pour le réseau de prédiction, et T prises pour le réseau de score.

$$\begin{aligned}\mathcal{L}_{cls}(a,p) = -\frac{1}{4P}\sum_{i=1}^{4P} AM(a_i)log(p_i) + \\ (1 - AM(a_i))log(1 - p_i)\end{aligned} \tag{4.3}$$

$$\begin{aligned}\mathcal{L}_{score}(g,p_s) = -\frac{1}{T}\sum_{i=1}^{T} JM(g_i)log(p_{si}) + \\ (1 - JM(g_i))log(1 - p_{si})\end{aligned} \tag{4.4}$$

où a_i et l'ancre considérée pour la $i^{ème}$ prédiction, g_i est la $i^{ème}$ prise prédite et p_i (respectivement p_{si}) est la probabilité de la prise produite par le réseau de prédiction (respectivement le réseau de score) et $AM(a_i)$ et $JM(g_i)$) valent 1 si la $i^{ème}$ prédiction est proche de la valeur réelle et 0 sinon.

La perte de régression est composée de deux termes : une fonction smooth L1 classique pour assurer que les prises prédites correspondent aux prises réelles et un second terme impliquant le réseau de score.

$$\begin{aligned}\mathcal{L}_{reg}(g,p_s) = \frac{\alpha}{P}\sum_{i=1}^{P}\sum_{m\in\{x,y,w,h,\theta\}} L1_{smooth}(\delta_{mi} - \hat{\delta_{mi}}) \\ - \frac{1}{T}\sum_{i=1}^{T} log(p_{si})\end{aligned} \tag{4.5}$$

où δ_i sont les valeurs de régression prédites et $\hat{\delta}_i$ sont les valeurs réelles. α est un hyperparamètre (fixé à 2 dans nos expérimentations).

Grâce à cette fonction de perte, le réseau de prédiction de prise sera capable d'utiliser le réseau de score pour estimer la qualité de ses prédictions et de les modifier pour améliorer leur qualité. Durant la rétropropagation, les gradients de cette perte sont uniquement utilisés pour mettre à jour les poids de la couche de régression et de l'extracteur de descripteurs, mais pas le réseau de score. De la même manière, les gradients de $\mathcal{L}_{score}$ ne sont pas utilisés pour mettre à jour la couche de régression, mais seulement le réseau de score et l'extracteur de descripteurs.

Un des jeux de données les plus utilisés pour entraîner et évaluer des réseaux de détection de prises est Cornell Grasping Dataset. Cependant, les modèles de l'état de l'art obtiennent maintenant des performances très élevées sur ces données. Les quelques erreurs de détection correspondent en réalité à des prises cohérentes mais qui ne correspondent pas à des prises annotées. Nous avons donc décidé d'utiliser une base plus réaliste et plus compliquée : la base Jacquard que nous avons proposée

dans [DDC18]. Les données ont été divisées en 5 parties afin d'utiliser une validation croisée.

Les résultats sont présentés dans la Figure 4.10. Le critère de performance est le Jaccard Matching (une prise est considérée bonne si elle est proche de la prise réelle (aussi bien en orientation qu'en intersection sur l'union) avec des valeur seuil de 30° pour l'angle et 25% pour l'intersection sur l'union. Quelques exemples de prises sont par ailleurs montrés dans la Figure 4.9.

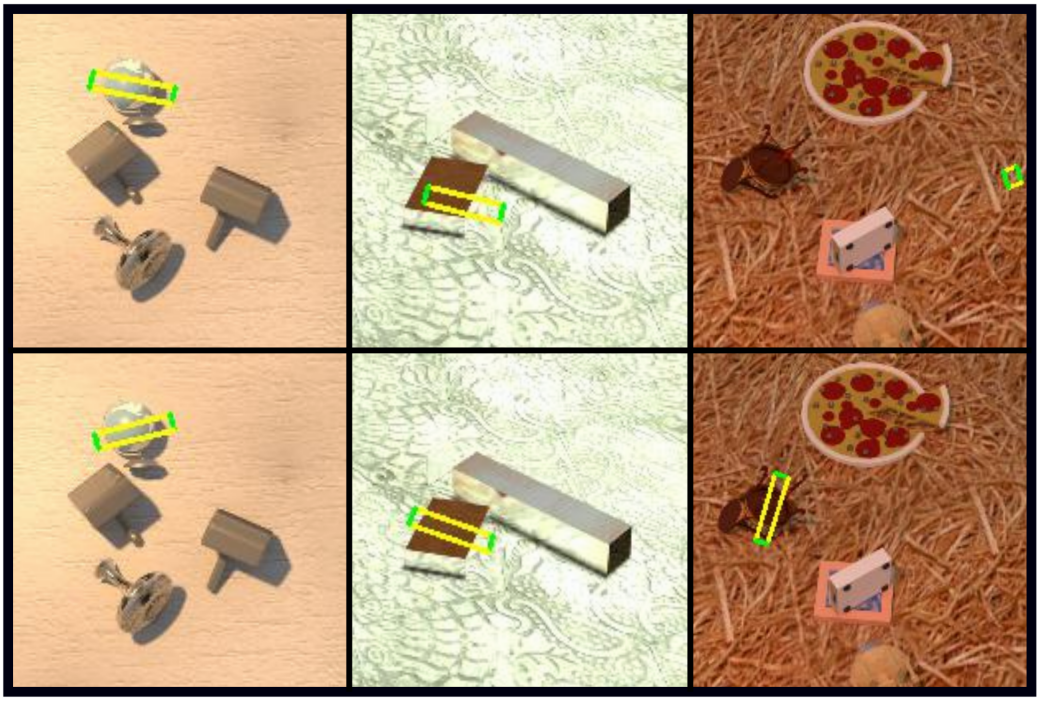


FIGURE 4.9 – Comparaison de propositions de prises en utilisant l'architecture de [ZLZ+18] (ligne supérieure), et la nôtre (ligne inférieure).

Comme on peut l'observer entre les lignes 2 et 3, la précision du réseau de score est meilleure que celle de la baseline, passant de 81,95% à 82,36% pour les prises en top 1 et de 72,07% 75,80% pour les prises en top 10. Même si l'augmentation n'est pas très importante, cela montre cependant que le réseau de score est capable de fournir des prédictions précises de prises, même sans utiliser les scores intermédiaires pour l'apprentissage. La deuxième et la quatrième ligne représente la précision en utilisant les prédictions ordonnées avec les scores intermédiaires, sans le réseau de score pour l'apprentissage pour la deuxième ligne, et avec pour la quatrième. Utiliser le réseau de score pendant l'apprentissage permet donc d'améliorer la qualité de prédiction des prises. Des deux dernières lignes, nous pouvons conclure que les scores prédits par le réseau de score sont plus précis que ceux fournis par la couche de score intermédiaire puisque trier les mêmes prédictions de prises en utilisant les valeurs

	Jacquard dataset accuracy		
Architecture	**Top 1 grasp**	**Top 5 grasps**	**Top 10 grasps**
Depierre *et al.*	74.21%	-	-
Zhou *et al.*	81.95%	77.97%	72.07%
Ours, without intermediate score	82.36%	79.73%	75.80%
Ours (intermediate score output)	83.61%	79.40%	73.87%
Ours (scorer output)	**85.74%**	**82.96%**	**79.37%**

FIGURE 4.10 – Comparaison des performances de différentes méthodes pour la prédiction de prises sur le jeu de données Jacquard. Références : Depierre et al. [DDC18], Zhou et al. [ZLZ+18].

prédites par le réseau de score permet d'obtenir des prises de meilleure qualité par rapport à l'utilisation des scores intermédiaires. Ainsi, l'architecture que nous avons proposée est plus efficace pour la prédiction de prises que les solutions de l'état de l'art.

4.4 Conclusion

Nous avons présenté dans ce chapitre nos principales contributions dans le domaine de la vision par ordinateur et l'apprentissage automatique appliquées à la robotique. L'application visée est le «picking», c'est à dire doter un bras robotique de la capacité à saisir des objets dans un vrac.

Nous avons tout d'abord proposé une approche afin d'identifier dans un vrac dense et homogène les instances d'objets les plus prenables. Ainsi, notre méthode permet de détecter les contours et les cotés occultants des objets à partir d'images RGB, en se basant sur une architecture encodeur-décodeur résiduel profond. Nous avons proposé d'augmenter la complexité du décodeur en couplant trois unités de type décodeurs légers et une unité de type encodeur-décodeur, disposés en cascade. Cela permet de structurer le processus de décodage et la réutilisation d'informations spécifiques aux sous-tâches : localisation du contour, identification du côté occultant du contour, segmentation d'instances non-occultées. Les expérimentations ont été conduites sur le jeu de données Mikado que nous avons également proposé afin de disposer d'un nombre très important d'images de vracs contenant de nom-

breuses instances. Les résultats ont montré que l'architecture du décodeur que nous avons proposé permet un gain significatif de performance par rapport aux schémas encodeur-décodeur traditionnels.

De plus, pour le problème de détection de prises d'objets à partir d'image RGB, nous avons également élaboré une architecture à base de réseaux profonds combinant la régression des paramètres de prise avec l'évaluation de la qualité de la prise, et dont l'apprentissage combiné permet l'utilisation de l'estimation de la qualité pour améliorer la régression. Afin de permettre l'apprentissage de modèles profonds, le jeu de données Jacquard a été constitué, et utilisé dans nos expérimentations. Celles-ci ont montré que notre solution est plus performante que les méthodes de l'état de l'art pour la prédiction de prises.

Conclusion générale et perspectives

Ce mémoire retrace mes principales contributions dans le domaine de la compréhension automatique de scènes visuelles, depuis mon arrivée en 2004 à l'Ecole Centrale de Lyon comme Maître de Conférences. Plus précisément, ces travaux ont porté sur les thématiques de classification d'images, détection d'objets dans les images, prédiction de l'impact émotionnel de vidéos ou encore prédiction de prises d'objets pour des bras robotiques. Ils sont le résultat de collaborations avec plusieurs doctorants, et s'inscrivent dans plusieurs projets et relations partenariales. Une courte synthèse est donnée ci-dessous.

Synthèse des contributions

Classification d'images

Descripteurs basés sur une représentation parcimonieuse des images

L'objectif d'une représentation parcimonieuse est d'obtenir une représentation fidèle d'un signal pouvant être considéré comme une combinaison linéaire d'atomes constituant un dictionnaire de dimension très supérieure à celle du signal lui-même. Cette décomposition va introduire dans la nouvelle représentation du signal un grand nombre de valeurs nulles. Elle a été originellement proposée dans le domaine du traitement du signal comme un outil puissant pour acquérir, représenter et compresser des signaux de grande dimension. Des études ont également montré que ces principes s'appliqueraient aux neurones du cortex visuel qui utiliseraient un codage parcimonieux pour représenter efficacement des scènes naturelles. Ces intéressantes propriétés nous ont conduit à proposer une adaptation de ces principes au problème de la classification d'images.

Dans ce cadre, nous avons développé une approche reconstructive et discriminative (RD_SROC) s'appuyant sur la représentation parcimonieuse des images. Elle repose sur l'hypothèse intuitive que l'image peut être représentée par une combinaison linéaire des images d'apprentissage de la même catégorie. Par conséquent,

les représentations parcimonieuses des images sont d'abord calculées par la résolution du problème de minimisation de la norme $L1$ et sont ensuite utilisées en tant que nouveaux descripteurs pour les images afin de permettre la classification de ces dernières par des classifieurs traditionnels tels que SVM. Afin d'améliorer la capacité de discrimination de la représentation parcimonieuse pour mieux répondre au problème de classification, nous avons également inclus un terme de discrimination, comme la mesure de discrimination Fisher ou la sortie d'un classifieur SVM, à la fonction d'objectif de la représentation parcimonieuse standard afin d'entraîner un dictionnaire reconstructif et discriminatif. De plus, nous avons proposé de combiner le dictionnaire reconstructif et discriminatif avec le dictionnaire adapté purement reconstructif pour une catégorie donnée de sorte que la capacité de discrimination puisse être augmentée.

Descripteur textuel pour la caractérisation des images

Afin de compléter les informations portées par les descripteurs visuels, nous avons proposé un nouveau descripteur textuel dédié au problème de la classification d'images. En effet, la plupart des photos publiées sur des sites de partage en ligne (Flickr, Facebook, ...) sont accompagnées d'une description textuelle sous la forme de mots-clés ou de légende. Ces descriptions constituent une riche source d'information sur la sémantique contenue dans les images et il est donc particulièrement intéressant de les considérer dans un système de classification d'images. Ainsi, nous avons élaboré des descripteurs HTC («Histograms of Textual Concepts») pour capturer les liens sémantiques entre les concepts. L'idée générale derrière HTC est de représenter un document textuel comme un histogramme de concepts textuels selon un dictionnaire (ou vocabulaire), pour lequel chaque valeur associée à un concept est l'accumulation de la contribution de chaque mot du texte pour ce concept, en fonction d'une mesure de distance sémantique. Plusieurs variantes de HTC ont été proposées qui se sont révélées être très efficaces. Inspirés par la démarche de l'analyse cepstrale de la parole, nous avons également développé Cepstral HTC pour capturer à la fois l'information de fréquence d'occurrence des mots (comme TF-IDF) et les liens sémantiques entre concepts fournis par HTC à partir des mots-clés associés aux images.

Fusion multimodale

Lorsque plusieurs sources d'information sont à disposition pour caractériser des données visuelles, il devient nécessaire de les combiner avec pour objectif d'en tirer

le meilleur parti pour les concepts visuels à reconnaître.

Nous avons donc élaboré une méthode de fusion (SWLF pour «Selective Weighted Later Fusion») afin de combiner efficacement différentes sources d'information pour le problème de la classification d'images. Cette approche de fusion est conçue pour sélectionner les meilleurs descripteurs et pondérer leur contribution pour chaque concept à reconnaître. SWLF s'est révélé être particulièrement efficace pour fusionner des modalités visuelles et textuelles, par rapport à des schémas de fusion standards. Dans la mesure où une fusion tardive au niveau des scores des classifieurs est reconnue pour être une manière simple et efficace pour combiner des descripteurs de nature différente, SWLF s'appuie sur deux idées simples. Premièrement, le score de classification à partir d'un type de descripteur (classifieur expert) doit être pondéré en fonction de sa qualité intrinsèque pour le problème de classification en question. Deuxièmement, dans le cadre d'un scénario multi-labels où plusieurs concepts visuels peuvent être attribuées à une même image, différents concepts visuels peuvent nécessiter différents types de descripteurs pour permettre leur reconnaissance de manière efficace. Ce modèle de fusion multimodale a été utilisé dans le cadre de notre participation au challenge «Photo Annotation» de ImageCLEF en 2012 et nous a permis d'obtenir la $1^{ère}$ place parmi 80 soumissions de 18 équipes.

Détection d'objets dans les images

Apprentissage faiblement supervisé de modèles à parties déformables

Nous avons proposé une amélioration de l'approche «Deformable Part-based Models» (DPM) faiblement supervisée, en insistant sur l'importance de la position et de la taille du filtre racine initial spécifique à la classe. Tout d'abord, un ensemble de candidats est calculé, ceux-ci représentant les positions possibles de l'objet pour le filtre racine initial, en se basant sur une mesure générique d'objectness (par region proposals) pour combiner les régions les plus saillantes et potentiellement de bonne qualité. Ensuite, nous avons proposé l'apprentissage du label des classes latentes de chaque candidat comme un problème de classification binaire, en entraînant des classifieurs spécifiques pour chaque catégorie afin de prédire si les candidat sont potentiellement des objets cible ou non. De plus, nous avons amélioré la détection en incorporant l'information contextuelle à partir des scores de classification de l'image. Enfin, nous avons élaboré une procédure de post-traitement permettant d'élargir et de contracter les régions fournies par le DPM afin de les adapter efficacement à la taille de l'objet, augmentant ainsi la précision finale de la détection.

Détection d'objets semi-supervisée basée sur le transfert de connaissances visuelles et sémantiques

Pour la seconde approche, nous avons étudié dans quelle mesure l'information tirée des objets similaires d'un point de vue visuel et sémantique pouvait être utilisée pour transformer un classifieur d'images en détecteur d'objets d'une manière semi-supervisée sur un large ensemble de données, pour lequel seul un sous-ensemble des catégories d'objets est annoté avec des boîtes englobantes nécessaires pour l'apprentissage des détecteurs. Nous avons proposé de transformer des classifieurs d'images basés sur des réseaux convolutionnels profonds (Deep CNN) en détecteurs d'objets en modélisant les différences entre les deux en considérant des catégories disposant à la fois de l'annotation au niveau de l'image globale et l'annotation au niveau des boîtes englobantes. Cette information de différence est ensuite transférée aux catégories sans annotation au niveau des boîtes englobantes, permettant ainsi la conversion de classifieurs d'images en détecteurs d'objets.

Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos

LIRIS-ACCEDE : une plateforme de données pour l'analyse du contenu émotionnel de vidéos

Cette plateforme, LIRIS-ACCEDE, contient un grand nombre de vidéos variées sous licence «Creative Commons» et pouvant donc être librement diffusées. Elle est constituée de deux types d'annotation : 10900 extraits vidéos d'une dizaine de secondes sont annotés globalement selon la valence et l'arousal, et 66 films (36 heures) sont annotés de manière continue (chaque seconde) selon la valence, l'arousal, ainsi que la peur. La qualité de cette base a été reconnue d'une part par plusieurs publications dans les conférences et journaux du domaine de l'informatique affective et d'autre part par son adoption comme données d'apprentissage et de test pour les tâches «Affective Impact of Movies» à MediaEval 2015, et «Emotional Impact of Movies» à MediaEval 2016, 2018 et 2019. Le nombre de téléchargements est actuellement de 439 (janvier 2020).

Modèle spatio-temporel profond pour la prédiction de l'impact émotionnel des vidéos

Afin d'estimer les émotions induites par les films, nous avons proposé plusieurs modèles, les plus performants reposant sur l'apprentissage profonds. L'un de ces

modèles intègre l'information temporelle car en effet, l'émotion ressentie lors du visionnage d'une scène d'un film dépend non seulement de la scène courante, mais également des scènes précédentes ainsi que des émotions ressenties précédemment. Ainsi, ce modèle est composé de deux réseaux de neurones convolutifs ajustés. L'un est dédié à la modalité visuelle et utilise en entrée des versions recadrées des principales images extraites des segments vidéos, alors que l'autre est dédié à la modalité audio utilisant en entrée un spectrogramme. Les activations de la dernière couche entièrement connectée de chaque réseau sont concaténées pour nourrir un réseau de neurones récurrent utilisant des neurones spécifiques appelés «Long-Short-Term Memory» qui permettent l'apprentissage des dépendances temporelles entre des segments vidéo successifs. La performance obtenue par le modèle est comparée à celle d'un modèle basique similaire à l'état de l'art et montre des résultats très prometteurs mais qui reflètent la complexité de telles tâches. En effet, la prédiction automatique des émotions induites par les films est donc toujours une tâche très difficile qui est loin d'être complètement résolue.

Analyse visuelle pour la robotique

Localisation d'instances d'objets dans un vrac

Notre objectif ici est de délimiter les instances d'objets dans un vrac et d'inférer leur dispositions spatiales à partir d'une unique image RGB, de manière à identifier dans un vrac les instances d'objets les plus prenables pour un bras robotique. Nous avons ainsi proposé un réseau profond composé d'un encodeur et d'un décodeur couplant trois unités de type décodeurs légers et une unité de type encodeur-décodeur, disposés en cascade. Cela permet de structurer le processus de décodage et la réutilisation d'informations spécifiques aux sous-tâches de localisation du contour, identification du côté occultant du contour et segmentation d'instances non-occultées. L'apprentissage de ce modèle, comme pour tous réseaux profonds, nécessite une quantité importante de données annotées. Or, produire une quantité suffisante d'images annotées manuellement avec les informations de contour et d'occlusions est inenvisageable. Nous avons donc également proposé un système pour générer des images synthétiques réalistes d'objet texturés disposés en vracs, Mikado, permettant un apprentissage efficace, et une bonne généralisation en situation avec un robot réel.

Prédiction de prises pour des bras robotiques

Au delà de la segmentation des instances les plus prenables, nous nous sommes intéressés justement à la prédiction des paramètres de prises sur les objets présents dans l'image. Ces paramètres définissent la position, l'orientation et l'ouverture de la pince utilisée pour saisir l'objet. Nous avons ainsi proposé une nouvelle architecture à base de réseaux profonds combinant la régression des paramètres de prise avec l'évaluation de la qualité de la prise, et dont l'apprentissage combiné permet l'utilisation de l'estimation de la qualité pour améliorer la régression. Afin de permettre un apprentissage et une évaluation efficace de ce modèle, nous avons également élaboré le jeu de de données **Jacquard** constituée de 54 485 scènes différentes à partir de 11 619 objets distincts avec un total de 4 967 454 annotations de prises. Celui-ci a actuellement été téléchargé par 53 équipes (janvier 2020).

Perspectives

Comme mentionné précédemment, le domaine de l'analyse de scènes visuelles a connu une rapide et profonde révolution avec l'avènement de modèles à base de réseaux de neurones convolutifs profonds. En effet, leur capacité à apprendre simultanément une représentation appropriée des images de haut niveau sémantique et à réaliser une classification ou une régression en fait des modèles extrêmement performants pour de nombreuses tâches (classification d'images, détection d'objets, reconnaissance de visages, segmentation sémantique, détection de prises pour des bras robotiques, ...).

Afin de permettre l'apprentissage de tels modèles, une quantité très importante de données est nécessaire. Cette problématique a rapidement suscité l'intérêt de la communauté de recherche et continue d'être un sujet très étudié car de nombreux verrous sont encore à lever.

Une solution consiste à développer des jeux de données très vastes, avec des données réelles lorsque cela est possible, comme nous l'avons fait avec la plateforme LIRIS-ACCEDE, ou des données de synthèse obtenues par simulation, ce que nous avons proposé avec la base Mikado pour la détection d'objets prenables par des bras robotiques, et Jacquard pour la détection de prises sur des objets pour des bras robotiques. Or dans ce dernier cas, il existe le risque que les données simulées soient trop éloignées des propriétés des données réelles et que donc le modèle appris sur des données de synthèse ne se généralise pas correctement aux données réelles.

Un autre solution consiste à intégrer ce problème de la famine des données dans

les stratégies d'apprentissage des modèles. C'est ce que nous avons proposé dans le cas de la détection d'objets dans les images avec notamment une stratégie de transfert de connaissances pour adapter des classifieurs d'images en détecteurs d'objets dans les images.

Pour aller plus loin, nous envisageons dans nos travaux futurs en particulier trois pistes qui nous semblent très prometteuses.

D'une part, pour réduire l'écart entre des données simulées et des données réelles, nous nous intéressons dans la thèse de Thomas Duboudin à une solution s'appuyant sur les réseaux antagonistes génératifs (GAN pour «Generative Adversarial Networks»), en particulier dans le cas de la détection d'objets en vue d'applications aéroportées et terrestres. En effet, l'objectif des GAN est de générer des données très réalistes. Pour cela, deux réseaux sont en compétition : un premier (le générateur) qui génère des données (images) et un second (le discriminateur) qui tente de déterminer s'il s'agit de données réelles ou produites par le générateur. L'apprentissage des deux réseaux se fait simultanément. Ces modèles font l'objet actuellement d'un nombre croissant de travaux de recherche, mais il persiste de nombreux problèmes, en particulier liés aux difficultés de convergence.

Une seconde direction que nous souhaitons emprunter concerne la problématique de «Few-shot learning» qui vise à permettre l'apprentissage d'un modèle profond avec très peu de données. Dans ce cadre, des approches de type «meta-learning» (apprendre à apprendre) ont permis d'obtenir des premiers résultats très prometteurs. Cependant, bien qu'adaptées pour un apprentissage efficace à partir de peu de données pour une tâche donnée, ces méthodes introduisent un nouveau problème car un nombre important de tâches doit être disponible pour permettre l'apprentissage d'une nouvelle tâche. De plus, les méthodes actuelles sont relativement efficaces lorsque les tâches d'apprentissage et de test sont relativement similaires, mais les performances chutent lorsque les tâches deviennent trop différentes. Enfin, aucune des approches actuelles ne considère la situation souvent rencontrée dans la réalité où les données ne sont disponibles qu'au fur et à mesure au cours du temps.

Enfin, pour les applications robotiques, et en particulier pour les tâches de picking/kitting, il est important de rendre possible la capitalisation de connaissances d'un robot ainsi que le transfert de connaissances entre robots pour permettre aux robots d'être flexibles, adaptables et autonomes dans des contextes instables et évolutifs. Nous envisageons donc de doter les robots d'une capacité de mémoire, à l'aide de réseaux profonds, afin qu'ils soient en mesure de capitaliser les connaissances acquises durant leurs expériences passées puis de les mutualiser par apprentissage par transfert.

Ces pistes sont très ambitieuses, mais des progrès auraient un impact très important sur la compréhension automatique de scènes visuelles, et plus largement dans les domaines de vision par ordinateur et apprentissage automatique.

Bibliographie

[ABC+18] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, and Saverio Blasi. Trecvid 2018 : Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[AEB06] M. Aharon, M. Elad, and A. Bruckstein. K-svd : An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11) :4311–4322, Nov 2006.

[AGM14] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 329–344, Cham, 2014. Springer International Publishing.

[AL12] Hossein Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *Computer Vision – ECCV 2012 : 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I*, number PART 1 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 836–849, 2012. QC 20121210.

[All77] J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3) :235–238, June 1977.

[AQG07] Stéphane Ayache, Georges Quénot, and Jérôme Gensel. Classifier fusion for svm-based multimedia semantic indexing. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 494–504, Berlin, Heidelberg, 2007. Springer-Verlag.

[Bav15] Yoann Baveye. *Automatic prediction of emotions induced by movies*. Theses, Ecole Centrale de Lyon, November 2015.

[BC61] R. Bellman and Karreman Mathematics Research Collection. *Adaptive Control Processes : A Guided Tour*. Princeton Legacy Library. Princeton University Press, 1961.

[BDCC15] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede : A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1) :43–55, Jan 2015.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[BL94] M. M. Bradley and P. J. Lang. Measuring emotion : The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1) :49–59, 1994.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 :993–1022, March 2003.

[BPT14] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised detection with posterior regularization. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[BPT15] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1081–1089, June 2015.

[CBD+90] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. Handwritten digit recognition with a back-propagation network. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[CBPZBA17] Souad Chaabouni, Jenny Benois-Pineau, Akka Zemmari, and Chokri Ben Amar. *Deep Saliency : Prediction of Interestingness in Video with CNN*, pages 43–74. Springer International Publishing, Cham, 2017.

[CDF+04] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[CL11] Chih-Chung Chang and Chih-Jen Lin. Libsvm : A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), May 2011.

[CSD+13] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton. Gtrace : General trace program compatible with emotionml. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 709–710, Sep. 2013.

[CV95] Corinna Cortes and Vladimir Vapnik. Support vector networks. *Machine Learning*, 20(3) :273–297, September 1995.

[CVS14] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2409–2416, June 2014.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing.

[DCB+16] E. Dellandrea, L. Chen, Y. Baveye, M. Sjöberg, and C. Chamaret. The mediaeval 2016 emotional impact of movies task. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, Hilversum, The Netherlands, October 20-21 2016.

[DDC18] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard : A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1-5, 2018*, pages 3511–3516, 2018.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet : A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[DF11] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR 2011*, pages 1777–1784, June 2011.

[DH08] M. Dikmen and T. S. Huang. Robust estimation of foreground in surveillance videos by sparse error estimation. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008.

[DHC+17] E. Dellandrea, M. Huigsloot, L. Chen, Y. Baveye, and M. Sjöberg. The mediaeval 2017 emotional impact of movies task. In *Working*

Notes Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, September 13-15 2017.

[DHC+18] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, Z. Xiao, and M. Sjöberg. The mediaeval 2018 emotional impact of movies task. In *Working Notes Proceedings of the MediaEval 2018 Workshop*, Sophia Antipolis, France, October 29-31 2018.

[DNR18] T. Do, A. Nguyen, and I. Reid. Affordancenet : An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5882–5889, May 2018.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[Dum04] Susan T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1) :188–230, 2004.

[EA06] M. Elad and M. Aharon. Image denoising via learned dictionaries and sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 895–900, June 2006.

[EAH99] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 5, pages 2443–2446 vol.5, March 1999.

[EEVG+15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge : A retrospective. *International Journal of Computer Vision*, 111(1) :98–136, January 2015.

[EGW+10] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2) :303–338, June 2010.

[Ekm99] Paul Ekman. Basic emotions. In Tim Dalgleish and M. J. Powers, editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley, 1999.

[FDC11] Huanzhang Fu, Emmanuel Dellandréa, and Liming Chen. Reconstructive and discriminative sparse representation for visual object

categorization. *Proceedings of the British Machine Vision Conference (BMVC)*, 01 2011.

[FE08] Jonathan M. Fishbein and Chris Eliasmith. Integrating Structure and Meaning : A New Method for Encoding Structure for Text Classification. In *Proceedings of the 30th European Conference on Information Retrieval Research : Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 514–521. Springer International Publishing, 2008.

[Fel98] Christiane Fellbaum, editor. *WordNet : an electronic lexical database*. MIT Press, 1998.

[FGMR10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9) :1627–1645, Sep. 2010.

[FKH+19] P. Follmann, R. König, P. Härtinger, M. Klostermann, and T. Böttger. Learning to see the invisible : End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336, Jan 2019.

[Fle71] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5) :378–382, 1971.

[FZD+09] H. Fu, C. Zhu, E. Dellandréa, C. Bichot, and L. Chen. Visual object categorization via sparse representation. In *Fifth International Conference on Image and Graphics*, pages 943–948, Sep. 2009.

[GCL89] M. K. Greenwald, E. W. Cook, and P. J. Lang. Affective judgment and psychophysiological response : Dimensional covaration in the evaluation of pictorial stimuli. *Journal of Psychophysiology*, 3(1) :51 – 64, 1989.

[GDBBP16] Iván González-Díaz, Vincent Buso, and Jenny Benois-Pineau. Perceptual modeling in the problem of active object recognition in visual scenes. *Pattern Recognition*, 56 :129 – 141, August 2016.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, page 580–587, USA, 2014. IEEE Computer Society.

[GFM11] Ross B. Girshick, Pedro F. Felzenszwalb, and David McAllester. Object detection with grammar models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 442–450, USA, 2011. Curran Associates Inc.

[GIDM15] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 437–446, June 2015.

[GL95] James J. Gross and Robert W. Levenson. Emotion elicitation using films. *Cognition and Emotion*, 9(1) :87–108, 1995.

[GLFM15] Syntyche Gbehounou, François Lecellier, and Christine Fernandez-Maloigne. Image database indexing : Emotional impact assessing. *Electronic Letters on Computer Vision and Image Analysis*, 14(3), December 2015.

[GLFM16] Syntyche Gbèhounou, François Lecellier, and Christine Fernandez-Maloigne. Evaluation of local and global descriptors for emotional impact recognition. *Journal of Visual Communication and Image Representation*, 38 :276 – 283, 2016.

[GR13] Yu Guo and Su Ruan. *Signal Separation with A Priori Knowledge Using Sparse Representation*, pages 315–332. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[Gra19] Matthieu Grard. *Generic instance segmentation for object-oriented bin-picking.* PhD thesis, Ecole Centrale de Lyon, 2019. Thèse de doctorat dirigée par Chen, Liming et Dellandréa, Emmanuel Informatique Lyon 2019.

[GS05] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5) :602 – 610, 2005. IJCNN 2005.

[HA07] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 609–616. MIT Press, 2007.

[Han06] A. Hanjalic. Extracting moods from pictures and sounds : towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2) :90–100, March 2006.

[HDMZ14] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 801–804, New York, NY, USA, 2014. Association for Computing Machinery.

[HGT+14] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda : Large scale detection through adaptation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3536–3544. Curran Associates, Inc., 2014.

[HL08] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 39–43, New York, NY, USA, 2008. ACM.

[HLvdMW17] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, July 2017.

[HMQ16] Abdelkader Hamadi, Philippe Mulhem, and Georges Quénot. A comparative study for multiple visual concepts detection in images and videos. *Multimedia Tools and Applications*, 75(15) :8973–8997, August 2016.

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8) :1735–1780, November 1997.

[HTL10] Mark J. Huiskes, Bart Thomee, and Michael S. Lew. New trends and ideas in visual concept detection : The mir flickr retrieval evaluation initiative. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, pages 527–536, New York, NY, USA, 2010. ACM.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[JWHY08] Jianchao Yang, J. Wright, T. Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[KAB+19] Mohammad Reza Kavoosifar, Daniele Apiletti, Elena Baralis, Paolo Garza, and Benoit Huet. Effective video hyperlinking by means of enriched feature sets and monomodal query combinations. *International Journal of Multimedia Information Retrieval, 10 June 2019*, 06 2019.

[Kri70] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1) :61–70, 1970.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[KTJ06] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 419–426, Washington, DC, USA, 2006. IEEE Computer Society.

[LBC99] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps) : Technical manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida*, 1999.

[LCHR19] Y. Liu, S. Canu, P. Honeine, and S. Ruan. Mixed integer programming for sparse coding : Application to image denoising. *IEEE Transactions on Computational Imaging*, 5(3) :354–365, Sep. 2019.

[LDC+13] Ningning Liu, Emmanuel Dellandréa, Liming Chen, Chao Zhu, Yu Zhang, Charles-Edmond Bichot, Stéphane Bres, and Bruno Tellez. Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. *Computer Vision and Image Understanding*, 117(5) :493 – 512, 2013.

[LDTC14] Ningning Liu, Emmanuel Dellandréa, Bruno Tellez, and Liming Chen. *A Selective Weighted Late Fusion for Visual Concept Recognition*, pages 1–28. Springer International Publishing, Cham, 2014.

[Lie98] Rainer W. Lienhart. Comparison of automatic shot boundary detection algorithms. In Minerva M. Yeung, Boon-Lock Yeo, and Charles A. Bouman, editors, *Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 290 – 301. International Society for Optics and Photonics, SPIE, 1998.

[LK77] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1) :159–174, 1977.

[LLM+18] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9628–9639, Red Hook, NY, USA, 2018. Curran Associates Inc.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco : Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[LMMX14] Martha Larson, Mark Melenhorst, María Menéndez, and Peng Xu. *Using Crowdsourcing to Capture Complexity in Human Interpretations of Multimedia Content*, pages 229–269. Springer International Publishing, Cham, 2014.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2) :91–110, November 2004.

[LSG+17] M. Larson, M. Soleymani, G. Gravier, B. Ionescu, and G. J. F. Jones. The benchmarking initiative for multimedia evaluation : Mediaeval 2016. *IEEE MultiMedia*, 24(1) :93–96, Jan 2017.

[LSZ04] Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations : An experimental comparison. In *International Conference on Information and Knowledge Management, Proceedings*, pages 615–624, 01 2004.

[MB04] Alessandro Moschitti and Roberto Basili. Complex linguistic features for text classification : A comprehensive study. In Sharon McDonald and John Tait, editors, *Advances in Information Retrieval*, pages 181–196, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[MB13] S. Marioryad and C. Busso. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 85–90, Sep. 2013.

[MBP+08] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[MCDC12] Henning Mller, Paul Clough, Thomas Deselaers, and Barbara Caputo. *ImageCLEF : Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 2012.

[Mil95] George A. Miller. Wordnet : A lexical database for english. *Commun. ACM*, 38(11) :39–41, November 1995.

[MN13] A. Metallinou and S. Narayanan. Annotation and processing of continuous emotional attributes : Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[MZ93] S. G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12) :3397–3415, Dec 1993.

[NALV18] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Semi-convolutional operators for instance segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[NGP11] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-

arousal space. *IEEE Transactions on Affective Computing*, 2(2) :92–105, April 2011.

[OF96] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381 :607–609, 1996.

[OF97] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set : A strategy employed by v1 ? *Vision Research*, 37(23) :3311 – 3325, 1997.

[OSL00] Bruno A. Olshausen, Phil Sallee, and Michael S. Lewicki. Learning sparse image codes using a wavelet pyramid architecture. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 851–857, Cambridge, MA, USA, 2000. MIT Press.

[PBAC+17] Stefano Pini, Olfa Ben Ahmed, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Benoit Huet. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *ICMI 2017, 19th ACM International Conference on Multimodal Interaction, November 13-17th, 2017, Glasgow, United Kingdom*, Glasgow, UNITED KINGDOM, 11 2017.

[PD07] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[PL11] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision*, pages 1307–1314, Nov 2011.

[PLU80] ROBERT PLUTCHIK. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3 – 33. Academic Press, 1980.

[PNK94] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11) :1119 – 1125, 1994.

[PRK93] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1, Nov 1993.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[QJL+19] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.

[Ran10] Justus Randolph. Free-marginal multirater kappa (multirater kfree) : An alternative to fleiss fixed-marginal multirater kappa. In *Advances in Data Analysis and Classification*, volume 4, 01 2010.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3) :211–252, 2015.

[REK+15] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66 :22 – 30, 2015. Pattern Recognition in Human Computer Interaction.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning Representations by Back-propagating Errors. *Nature*, 323(6088) :533–536, 1986.

[RM77] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3) :273 – 294, 1977.

[RR13] X. Ren and D. Ramanan. Histograms of sparse codes for object detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3246–3253, June 2013.

[RRG07] J. Rottenberg, R. D. Ray, and J. J. Gross, editors. *Emotion elicitation using films*. Handbook of emotion elicitation and assessment, Oxford University Press, 2007.

[RSS+10] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where – and why ? semantic relatedness for knowledge

transfer. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 910–917, June 2010.

[RTVY08] S. R. Rao, R. Tron, R. Vidal, and Yi Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[Rus03] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110 1 :145–72, 2003.

[RW15] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4315–4324, June 2015.

[SB14] J. Schlüter and S. Böck. Improved musical onset detection with convolutional neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983, May 2014.

[SBW+15] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The mediaeval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[SC04] Magnus Sahlgren and Rickard Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[SED05] J. . Starck, M. Elad, and D. L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14(10) :1570–1582, Oct 2005.

[SGJ+14] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1611–II–1619. JMLR.org, 2014.

[SK10] Scherer and Kr. The component process model : A blueprint for a comprehensive computational model of emotion. In Klaus R. Scherer,

Tanja Bänziger, and Etienne Roesch, editors, *A Blueprint for Affective Computing : A Sourcebook and Manual*. Oxford University Press, 2010.

[SLJD14] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 1637–1645, Cambridge, MA, USA, 2014. MIT Press.

[SLP02] Didier Schwab, Mathieu Lafourcade, and Violaine Prince. Antonymy and conceptual vectors. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[SP97] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11) :2673–2681, Nov 1997.

[SSX12] Zhiyuan Shi, Parthipan Siva, and Tony Xiang. Transfer learning by ranking for weakly supervised object annotation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2012.

[ST11] P. Siva and Tao Xiang. Weakly supervised object detector learning with model drift detection. In *2011 International Conference on Computer Vision*, pages 343–350, Nov 2011.

[SWS05] Cees Snoek, Marcel Worring, and Arnold Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th ACM International Conference on Multimedia, MM 2005*, pages 399–402, 01 2005.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11) :613–620, November 1975.

[SWY+15] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[TLN+03] B. L. Tseng, C. . Lin, M. Naphade, A. Natsev, and J. R. Smith. Normalized classifier fusion for semantic visual concept detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 2, pages II–535, Sep. 2003.

[TSBB+14] S. Tiberius Strat, A. Benoit, Hervé Bredin, Georges Quénot, and P. Lambert. Hierarchical Late Fusion for Concept Detection in Videos. In *Fusion in Computer Vision : Understanding Complex Visual Content*, pages 53–78. Springer international publishing, 2014.

[TWDC17] Y. Tang, X. Wang, E. Dellandréa, and L. Chen. Weakly supervised learning of deformable part-based models for object detection via region proposals. *IEEE Transactions on Multimedia*, 19(2) :393–407, Feb 2017.

[TWW+18] Y. Tang, J. Wang, X. Wang, B. Gao, E. Dellandréa, R. Gaizauskas, and L. Chen. Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12) :3045–3058, Dec 2018.

[USGS13] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int. J. Comput. Vision*, 104(2) :154–171, September 2013.

[Vap95] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

[WCCS04] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 572–579, New York, NY, USA, 2004. ACM.

[WES14] F. Weninger, F. Eyben, and B. Schuller. On-line continuous-time music mood regression with deep recurrent neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5412–5416, May 2014.

[WHR+15] Chong Wang, Kaiqi Huang, Weiqiang Ren, Junge Zhang, and Stephen J. Maybank. Large-scale weakly supervised object localization via latent category learning. *IEEE Transactions on Image Processing*, 24 :1371–1385, 2015.

[WJW01] J. Z. Wang, Jia Li, and G. Wiederhold. Simplicity : semantics-sensitive integrated matching for picture libraries. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence, 23(9) :947–963, Sep. 2001.

[WYG+09] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2) :210–227, Feb 2009.

[YC11] Y. Yang and H. H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4) :762–774, May 2011.

[ZF14] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[ZLZ+18] Xinwen Zhou, Xuguang Lan, Hanbo Zhang, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Fully convolutional grasp detection network with oriented anchor box. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7223–7230, 2018.

[ZMLS07] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories : A comprehensive study. *International Journal of Computer Vision*, 73(2) :213–238, Jun 2007.

[ZTMD17] Yan Zhu, Yuandong Tian, Dimitris N. Metaxas, and Piotr Dollár. Semantic amodal segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3009, 2017.

Rapport d'activités

Informations personnelles

42 ans, né le 10 mars 1977 à Thionville, France
Marié, 2 enfants
Maître de Conférences en Informatique
Ecole Centrale de Lyon
Laboratoire LIRIS (UMR5205)
Section CNU 27
Bénéficiaire de la **PES** puis de la **PEDR** depuis 2013

Adresse : 36 avenue Guy de Collongue, 69134 Ecully Cedex
Téléphone : 04 72 18 65 26
Adresse électronique : emmanuel.dellandrea@ec-lyon.fr
Site internet : http ://perso.ec-lyon.fr/emmanuel.dellandrea/

Formation universitaire

- 2000-2003 : **Doctorat en Informatique** de l'Université de Tours.
  - *Titre :* Analyse de signaux vidéos et sonores : application à l'étude de signaux médicaux.
  - *Directeurs :* Pr. Nicole Vincent et Dr. Pascal Makris
- 1999-2000 : **DEA Signaux et Images en Biologie et Médecine** de l'Université de Tours.
  - *Titre du stage :* Suivi d'objets déformables dans une séquence d'images : application à la caractérisation du sphincter œsophagien inférieur.
  - *Directeurs :* Pr. Nicole Vincent et Dr. Pascal Makris
- 1997-2000 : **Diplôme d'Ingénieur en Informatique** de l'Ecole Polytechnique de l'Université de Tours.

Parcours professionnel

- 2004-présent : **Maître de Conférences** à l'Ecole Centrale de Lyon, membre du laboratoire LIRIS.
- 2003-2004 : **Attaché Temporaire d'Enseignement et de Recherche** au Département Informatique de l'Ecole Polytechnique de l'Université de Tours.
- 2000-2003 : **Allocataire de recherche et moniteur en Informatique** à l'Université de Tours.

Encadrement doctoral

Depuis mon arrivée à l'Ecole Centrale de Lyon en 2004, j'ai co-encadré et co-encadre actuellement les 10 doctorants suivants :

Thomas Duboudin [2019-présent] : Apprentissage profond de simulation augmentée en vue d'applications aéroportées et terrestres (financement CIFRE avec Thalès, co-encadrement avec Pr. Liming Chen)

Amaury Depierre [2017-présent] : Deep Reinforcement Learning for Adaptative Robotic Grasping (financement CIFRE avec Siléane, co-encadrement avec Pr. Liming Chen)

Matthieu Grard [2015-2019] : Generic Instance Segmentation for Object-Oriented Bin-Picking (financement CIFRE avec Siléane, co-encadrement avec Pr. Liming Chen)

Yuxing Tang [2010-2016] : Weakly Supervised Learning of Deformable Part Models and Convolutional Neural Networks for Object Detection (financement Chinese Scolarship Council, co-encadrement avec Pr. Liming Chen)

Yoann Baveye [2012-2015] : Automatic prediction of emotions induced by movies (financement CIFRE avec Technicolor, co-encadrement avec Dr. Christel Chamaret et Pr. Liming Chen)

Ningning Liu [2009-2013] : Contributions to generic and affective visual concept recognition (financement Chinese Scolarship Council, co-encadrement avec Dr. Bruno Tellez et Pr. Liming Chen)

Boyang Gao [2009-2014] : Contributions to music semantic analysis and its acceleration techniques (financement Chinese Scolarship Council, co-encadrement avec Pr. Liming Chen)

Huanzhang Fu [2006-2010] : Contributions to generic visual object categorization (financement Chinese Scolarship Council, co-encadrement avec Pr. Liming Chen)

Xi Zhao [2007-2010] : 3D face analysis : landmarking, expression recognition and beyond (financement Chinese Scolarship Council, co-encadrement avec Pr. Liming Chen)

Zhongzhe Xiao [2004-2008] : Recognition of emotions in audio signals (financement Chinese Scolarship Council founding, co-encadrement avec Pr. Weibei Dou et Pr. Liming Chen)

Projets de recherche

Participation à des projets de recherche

Je participe actuellement aux projets :

CHIST-ERA Learn-Real [2019-2022] : Amélioration de la reproductibilité dans l'apprentissage de compétences de manipulations physiques par des robots en utilisant des simulateurs avec des variations réalistes. Partenaires : Idiap Research Institute (Suisse) et Fondazione Istituto Italiano di Tecnologia (Italie).

LABCOM Arès [2017-2020] : Apprentissage et Vision par Ordinateur pour robots intelligents. Partenaire : Siléane.

FUI Pikaflex [2016-2020] : Développement de solutions de Picking/Kitting automatisées et flexibles pour des bras robotiques industriels. Partenaires : Renault, Siléane.

Par ailleurs, j'ai auparavant participé aux projets :

ANR VideoSense [2010-2013] : Annotation automatique de vidéos avec des concepts de haut-niveau sémantique. Partenaires : EURECOM, LIG, LIF et GHANNI.

CNRS PICS MusicFinder [2007-2009] : Développement de techniques d'analyse de titres musicaux. Partenaire : Département de Génie Electrique de l'Université Tsinghua à Pékin.

ACI Musicdiscover [2004-2007] : Développement de méthodes pour la description sémantique structurée de la musique. Partenaires : IRCAM et LTCI.

Responsabilité scientifique de projets de recherche

J'ai été le responsable scientifique pour le LIRIS des deux projets suivants :

CHIST-ERA Visen [2013-2016] : Génération automatique de texte pour décrire le contenu d'images. Partenaires : Université de Surrey (Royaume-Uni), Université de Sheffield (Royaume-Uni) et IRI (Espagne).

ANR Omnia [2008-2010] : Filtrage automatique de documents contenant des images et du texte. Partenaires : LIG et Xerox XRCE.

Participation à des compétitions/challenges

Nous avons participé en 2011 et 2012 à la compétition internationale **ImageCLEF** dans le cadre de la tâche *Photo annotation* dont le but est l'annotation automatique d'un grand nombre d'images selon une centaine de concepts.

Notre soumission LIRIS a obtenu la seconde meilleure performance en 2011 parmi les 79 soumissions réalisées par 18 équipes internationales, et la meilleure performance en 2012 parmi les 80 soumissions réalisées par 18 équipes internationales.

Animation de la recherche

Participation à des jurys de thèse

J'ai participé en tant qu'examinateur au jury de thèse de Syntyche Gbehounou, dont le mémoire s'intitule «Indexation de bases d'images : Evaluation de l'impact émotionnel». La soutenance s'est déroulée le 21 novembre 2014 à l'Université de Poitiers.

Organisation d'évènements

J'ai été membre du comité d'organisation de la conférence CORESA qui s'est déroulée à Lyon en octobre 2010.

J'ai également organisé avec deux collègues la journée scientifique «Emotion» du GDR ISIS qui s'est déroulée en novembre 2010, puis j'ai été l'organisateur principal de sa deuxième édition en octobre 2013 «Reconnaissance, analyse et interprétation des émotions» dans le cadre du projet régional LIMA2, soutenue par le pôle de compétitivité Imaginove et le GdR ISIS (action «Visage, geste, action et comportement»). Son but était d'offrir l'opportunité d'échanges sous forme d'exposés et de discussions entre des chercheurs de différentes communautés (informaticiens, cogniticiens, psychologues, physiologues) travaillant sur l'émotion à partir de différents media et/ou modalités (son, image, vidéo, visage).

Enfin, j'ai participé à l'organisation de plusieurs compétitions internationales :

Emotional Impact of Movies à MediaEval : J'ai été l'organisateur principal de la compétition **Emotional Impact of Movies** en 2016 [A4], 2017 [A3] et 2018 [A2] dont l'objectif pour les participants était de développer des méthodes permettant de prédire l'impact émotionnel que le contenu de vidéos pourrait avoir sur les spectateurs.

Affective Impact of Movies à MediaEval 2015 [A5], portant sur le développent de systèmes pour détecter des contenus vidéo violents et prédire l'impact affectif des vidéos sur les spectateurs.

Scalable Image Annotation, Localization and Sentence Generation task à ImageCLEF en 2015 [A6], portant sur le développement de systèmes permettant de décrire les images, localiser les différents objets et générer une description de la scène.

HARL à ICPR 2012 [R8], portant sur la reconnaissance d'activités humaines à partir de vidéos.

Activités de relecture

J'exerce régulièrement des activités de relecture pour plusieurs revues et conférences internationales. Il s'agit des **revues internationales** suivantes : EURASIP Journal on Audio, Speech and Music Processing, EURASIP Journal on Advances in Signal Processing, EURASIP Journal on Image and Video Processing, Multimedia Tools and Applications, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Affective Computing, IEEE Transactions on Multimedia,

et des **conférences internationales** suivantes : International Conference on Image and Graphics (ICIG), International Conference on Computer Vision Theory and Application (VISAPP), Compression et Représentation des Signaux Audiovisuels (CORESA), Workshop on 3D Face Biometrics associé à la conférence IEEE International Conference on Automatic Face and Gesture Recognition (FG), IEEE International Conference on Image Processing (ICIP), International Conference on Affective Computing & Intelligent Interaction (ACII), International Conference on Multimedia Retrieval (ICMR), International Conference on Pattern Recognition (ICPR).

Par ailleurs, j'ai rapporté deux projets ANR en 2013 et un projet en 2014.

Autres responsabilités

Je suis **responsable de l'Unité d'Enseignement de l'Informatique** à l'Ecole Centrale de Lyon depuis 2010, ainsi que **responsable de l'équipe d'enseignement de l'Informatique** depuis 2018.

Résumé des activités d'enseignement

Je réalise mes activités d'enseignement à l'Ecole Centrale de Lyon où j'interviens dans les trois années du cycle d'ingénieur. J'enseigne à la fois des matières fondamentales de l'Informatique telles que l'**algorithmique**, la **programmation orientée objet**, la **programmation web**, ainsi que des matières étroitement liées à mes activités de recherche telles que l'**analyse de données**, la **reconnaissance de formes**, l'**indexation vidéo**, l'**analyse multimédia**, le **machine learning** et le **deep learning**.

Pour les années 2018-2019 et 2019-2020, j'assure également une intervention intitulée «Modèles computationnels pour la prédiction de l'émotion» dans le cadre du cours «Les émotions en activités» proposé par le Collège des Hautes Etudes Lyon Science[s] (CHELS) [2].

Liste des publications

Revues Internationales avec comité de lecture

[R1] Y. Tang, J. Wang, X. Wang, B. Gao, E. Dellandréa, R. Gaizauskas, L. Chen, "Visual and Semantic Knowledge Transfer for Large Scale Semi-supervised Object Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40(12), pp. 3045-3058, 2018.

[R2] Y. Lu, L. Chen, A. Saidi, E. Dellandréa, Y. Wang, "Discriminative Transfer Learning Using Similarities and Dissimilarities", IEEE Transactions on Neural Networks and Learning Systems, vol. 29(7), pp. 3097-3110, 2018.

[R3] Y. Baveye, C. Chamaret, E. Dellandréa, L. Chen, "Affective Video Content Analysis : A Multidisciplinary Insight", IEEE Transactions on Affective Computing, vol. 9(4), pp. 396-409, 2018.

[R4] Y. Tang, X. Wang, E. Dellandréa, L. Chen, "Weakly Supervised Learning of

2. www.chels.fr

Deformable Part-Based Models for Object Detection via Region Proposals", IEEE Transactions on Multimedia, vol. 19(2), pp. 393-407, 2017.

[R5] N. Liu, K. Wang, X. Jin, B. Gao, E. Dellandréa, L. Chen, "Visual affective classification by combining visual and text features", PloS one, vol. 12(8), 2017.

[R6] X. Zhao, J. Zou, H. Li, E. Dellandréa, I.-A. Kakadiaris, L. Chen, "Automatic 2.5-D Facial Landmarking and Emotion Annotation for Social Interaction Assistance", IEEE Transactions on Cybernetics, vol. 46(9), pp. 2042-2055, 2016.

[R7] Y. Baveye, E. Dellandréa, C. Chamaret, L. Chen, "LIRIS-ACCEDE : A Video Database for Affective Content Analysis", IEEE Transactions on Affective Computing, vol. 6(1), pp. 43-55, 2015.

[R8] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements", Computer Vision and Image Understanding, vol. 127, pp. 14-30, 2014.

[R9] X. Zhao, E. Dellandréa, J. Zou, L. Chen, "A unified probabilistic framework for automatic 3D facial expression analysis based on a Bayesian belief inference and statistical feature models", Image and Vision Computing, vol. 31(2), pp. 231-245, 2013.

[R10] N. Liu, E. Dellandréa, L. Chen, C. Zhu, Y. Zhang, C.-E. Bichot, S. Bres, B. Tellez, "Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme", Computer Vision and Image Understanding, vol. 117(5), pp. 493-512, 2013.

[R11] X. Zhao, E. Dellandréa, L. Chen, I.-A. Kakadiaris, "Accurate Landmarking of Three-Dimensional Facial Data in the Presence of Facial Expressions and Occlusions Using a Three-Dimensional Statistical Facial Feature Model", IEEE Transactions on Systems, Man and Cybernetics - Part B : Cybernetics, vol. 41(5), pp. 1417-1428, 2011.

[R12] C. Zhu, H. Fu, C.-E. Bichot, E. Dellandréa, L. Chen, "Visual object recognition using multi-scale local binary patterns and line segment feature", International Journal of Signal and Imaging Systems Engineering, 2011.

[R13] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Classification of Emotional Speech Based on an Automatically Elaborated Hierarchical Classifier", ISRN Signal Processing, 2011.

[R14] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Multi-stage Classification of Emotional Speech Motivated by a Dimensional Emotion Model", Multimedia Tools and Applications, vol. 46(1), pp. 119-145, 2010.

[R15] E. Dellandréa, P. Makris, N. Vincent, "Zipf Analysis of Audio Signals", Fractals, World Scientific Publishing Company, vol. 12(1), pp. 73-85, 2004.

Brevets

[B1] Y. Baveye, C. Chamaret, E. Dellandréa, L. Chen, "Methods for determining a personalized profile for filtering excerpts of a multimedia content, corresponding devices, computer program product and computer-readable carrier medium", Patent 16305106.3, 2017.

[B2] Y. Baveye, E. Dellandréa, C. Chamaret, L. Chen, "Method and apparatus for detecting emotional key frame", Patent 14306894.8, 2016.

Chapitres de livres

[L1] N. Liu, E. Dellandréa, B. Tellez, L. Chen, "A Selective Weighted Late Fusion for Visual Concept Recognition", Fusion in Computer Vision - Understanding Complex Visual Content, Springer International Publishing, pp. 1-28, 2014.

Conférences internationales avec comité de lecture

[C1] A. Depierre, E. Dellandréa, L. Chen, "Jacquard : A Large Scale Dataset for Robotic Grasp Detection", IEEE International Conference on Intelligent Robots and Systems (IROS), 2018.

[C2] M. Petit, A. Depierre, X. Wang, E. Dellandréa, L. Chen, "Developmental Bayesian Optimization of Black-Box with Visual Similarity-Based Transfer Learning", IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2018.

[C3] M. Grard, R. Brégier, F. Sella, E. Dellandréa, L. Chen, "Object segmentation in depth maps with one user click and a synthetically trained fully convolutional network", International Workshop on Human-Friendly Robotics (HFR), 2017.

[C4] Y. Tang, J. Wang, B. Gao, E. Dellandréa, R.-J. Gaizauskas, L. Chen, "Large Scale Semi-Supervised Object Detection Using Visual and Semantic Knowledge Transfer", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2119-2128, 2016.

[C5] Y. Baveye, E. Dellandréa, C. Chamaret, L. Chen, "Deep learning vs. kernel methods : Performance for emotion prediction in videos", International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 77-83, 2015.

[C6] A. Ramisa, J. Wang, Y. Lu, E. Dellandréa, F. Moreno-Noguer, R. Gaizauskas, "Combining Geometric, Textual and Visual Features for Predicting Prepositions in Image Descriptions", Conference on Empirical Methods in Natural Language Processing, pp. 214-220, 2015.

[C7] Y. Tang, X. Wang, E. Dellandréa, S. Masnou, L. Chen, "Fusing generic objectness and deformable part-based models for weakly supervised object detection", IEEE International Conference on Image Processing (ICIP), pp. 4072-4076, 2014.

[C8] Y. Baveye, E. Dellandréa, C. Chamaret, L. Chen, "From crowdsourced rankings to affective ratings", International Conference on Multimedia and Expo Workshops, pp. 1-6, 2014.

[C9] Y. Baveye, J.-N. Bettinelli, E. Dellandréa, L. Chen, C. Chamaret, "A Large Video Data Base for Computational Models of Induced Emotion",International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 13-18, 2013.

[C10] B. Gao, E. Dellandréa, L. Chen, "Sparse Music Decomposition onto a MIDI Dictionary Driven by Statistical Music Knowledge", International Society for Music Information Retrieval Conference (ISMIR), pp. 445-450, 2013.

[C11] B. Gao, E. Dellandréa, L. Chen, "Accelerated Dictionary Learning with GPU/Multicore CPU and its Application to Music Classification", International Conference on Signal Processing (ICSP), 2012.

[C12] H. Fu, E. Dellandréa, L. Chen, "Reconstructive and Discriminative Sparse Representation for Visual Object Categorization", British Machine Vision Conference (BMVC), 2011.

[C13] N. Liu, E. Dellandréa, B. Tellez, L. Chen, "Associating textual features with visual ones to improve affective image classification", International Conference on Affective Computing and Intelligent Interaction (ACII), 2011.

[C14] X. Zhao, E. Dellandréa, L. Chen, "Building a Statistical AU Space for Facial Expression Recognition in 3D", International Conference on Digital Image Computing : Techniques and Applications (DICTA), 2011.

[C15] N. Liu, E. Dellandréa, B. Tellez, L. Chen, "Evaluation of Features and Combination Approaches for the Classification of Emotional Semantics in Images", International Conference on Computer Vision, Theory and Applications (VISAPP), 2011.

[C16] S. Skaff, D. Rouquet, E. Dellandréa, A. Fallaise, V. Bellynck, H. Blanchon, C. Boitet, D. Schwab, L. Chen, A. Saidi, G. Csurka, L. Marchesotti, "Multimodal search for graphic designers", International Conference on Information Visualization Theory and Applications (IVAPP), 2011.

[C17] S. Wirkert, E. Dellandréa, L. Chen, "Bayesian GOETHE Tracking", International Conference on Pattern Recognition (ICPR), 2010.

[C18] X. Zhao, Di Huang, E. Dellandréa, L. Chen, "Automatic 3D Facial Expression Recognition based on a Bayesian Belief Net and a Statistical Facial Feature Model", International Conference on Pattern Recognition (ICPR), 2010.

[C19] X. Zhao, E. Dellandréa, L. Chen, D. Samaras, "AU Recognition on 3D Faces Based On An Extended Statistical Facial Feature Model", IEEE Fourth International Conference on Biometrics : Theory, Applications and Systems (BTAS), 2010.

[C20] H. Fu, A. Pujol, E. Dellandréa, L. Chen, "Image Modeling Using Statistical Measures for Visual Object Categorization", International Conference on Image Processing Theory, Tools and Applications (IPTA), 2010.

[C21] C. Zhu, H. Fu, C.-E. Bichot, E. Dellandréa, L. Chen, "Visual Object Recognition using Local Binary Patterns and Segment-based Feature", International Conference on Image Processing Theory, Tools and Applications (IPTA), 2010.

[C22] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Recognition of emotions in speech by a hierarchical approach", International Conference on Affective Computing and Intelligent Interaction (ACII), 2009.

[C23] X. Zhao, E. Dellandréa, L. Chen, "A People Counting System based on Face Detection and Tracking in a Video", IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2009.

[C24] H. Fu, Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Image Categorization Using ESFS : A New Embedded Feature Selection Method Based on SFS", Advanced Concepts for Intelligent Vision Systems (ACIVS), 2009.

[C25] X. Zhao, E. Dellandréa, L. Chen, "A 3D statistical facial feature model and its application on locating facial landmarks", Advanced Concepts for Intelligent Vision Systems (ACIVS), 2009.

[C26] H. Fu, C. Zhu, E. Dellandréa, C.-E. Bichot, L. Chen, "Visual Object Categorization via Sparse Representation", International Conference on Image and Graphics (ICIG), 2009.

[C27] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Two-stage classification of emotional speech", International Conference on Digital Telecommunications (ICDT'06), pp. 32-32, 2006.

[C28] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Features extraction and selection for emotional speech classification", IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 411-416, 2005.

[C29] E. Dellandréa, H. Harb, L. Chen, "Zipf, neural networks and SVM for mu-

sical genre classification", Fifth IEEE International Symposium on Signal Processing and Information Technology, pp. 57-62, 2005.

[C30] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Features extraction and selection for emotional speech classification", IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 411-416, 2005.

[C31] E. Dellandréa, P. Makris, N. Vincent, "Inner structure computation for audio signal analysis", International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 140-145, 2003.

[C32] E. Dellandréa, P. Makris, N. Vincent, "Wavelets and Zipf Law for Audio Signal Analysis", International Symposium on Signal Processing and its Applications (ISSPA), pp. 483-486, 2003.

[C33] E. Dellandréa, P. Makris, M. Boiron, N. Vincent, "Multiresolution for the detection of xiphoidal sounds in noisy medical audio signals", EURASIP Conference focused on Video/Image Processing and Multimedia Communication (EC-VIP-MC), pp. 619-624, 2003.

[C34] E. Dellandréa, P. Makris, C. Melin, M. Boiron, N. Vincent, "On the Contribution of Zipf Analysis for the Characterization of Medical Audio Signals", International Conference on Image and Signal Processing (ICISP), pp. 439-445, 2003.

[C35] E. Dellandréa, P. Makris, M. Boiron, N. Vincent, "Xiphoidal Sounds Analysis by the way of Audio Signal Codings and Global Parameters Extraction", International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, pp. 9-12, 2003.

[C36] E. Dellandréa, P. Makris, M. Boiron, N. Vincent, "A medical acoustic signal analysis method based on Zipf law", IEEE International Conference on Digital Signal Processing (DSP), pp. 615-618, 2002.

Workshops internationaux avec comité de lecture

[W1] T. Li, Y. Baveye, C. Chamaret, E. Dellandréa, L. Chen, "Continuous Arousal Self-assessments Validation Using Real-time Physiological Responses", Workshop on Affect and Sentiment in Multimedia, ACM Multimedia, pp. 39-44, 2015.

[W2] Y. Baveye, C. Chamaret, E. Dellandréa, L. Chen, "A Protocol for Cross-Validating Large Crowdsourced Data : The Case of the LIRIS-ACCEDE Affective Video Dataset", Workshop on Crowdsourcing for Multimedia (CrowdMM), ACM Multimedia, pp. 3-8, 2014.

[W3] B. Gao, E. Dellandréa, L. Chen, "Music sparse decomposition onto a MIDI dictionary of musical words and its application to music mood classification", Inter-

national Workshop on Content-Based Multimedia Indexing (CBMI), 2012.

[W4] N. Liu, E. Dellandréa, C. Zhu, C.-E. Bichot, L. Chen, "A Selective Weighted Late Fusion for Visual Concept Recognition", ECCV 2012 Workshop on Information Fusion in Computer Vision for Concept Recognition, 2012.

[W5] E. Dellandréa, N. Liu, L. Chen, "Classification of affective semantics in images based on discrete and dimensional models of emotions", International Workshop on Content-Based Multimedia Indexing (CBMI), 2010.

[W6] X. Zhao, P. Szeptycki, E. Dellandréa, L. Chen, "Precise 2.5D Facial Landmarking via an Analysis by Synthesis approach", IEEE Workshop on Applications of Computer Vision (WACV), 2009.

[W7] H. Fu, A. Pujol, E. Dellandréa, L. Chen, "Region based visual object categorization using segment features and polynomial modeling", Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 277-286, 2008.

[W8] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "What is the best segment duration for music mood analysis ?", International Workshop on Content-Based Multimedia Indexing, pp. 17-24,2008.

[W9] Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Automatic hierarchical classification of emotional speech", Ninth IEEE International Symposium on Multimedia Workshops (ISMW), pp. 291-296, 2007.

[W10] E. Dellandréa, P. Makris, M. Boiron, N. Vincent, "Active Contours for Bolus Tracking in X-Ray Images Sequences", IAPR Workshop on Machine Vision Applications (MVA), pp. 303-306, 2000.

Autres publications

[A1] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, Z. Xiao and M. Sjöberg, "Predicting the Emotional Impact of Movies", ACM SIGMM Records, Issue 4, 2018.

[A2] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, Z. Xiao, M. Sjöberg, "The MediaEval 2018 Emotional Impact of Movies Task", Working Notes Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, October 29-31, 2018.

[A3] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, M. Sjöberg, "The MediaEval 2017 Emotional Impact of Movies Task", Working Notes Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017.

[A4] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, C. Chamaret, "The MediaEval 2016 Emotional Impact of Movies Task", Working Notes Proceedings of the

MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016.

[A5] M. Sjöberg, Y. Baveye, H. Wang, V.-L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, L. Chen, "The MediaEval 2015 Affective Impact of Movies Task", Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015.

[A6] A. Gilbert, L. Piras, J. Wang, F. Yan, E. Dellandréa, R. Gaizauskas, V. Mauricio, K. Mikolajczyk, "Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task", ImageCLEF, CLEF2015 Working Notes, Toulouse, France, 2015.

Développement de plateformes de données

Afin de dynamiser la recherche dans le domaine de l'informatique affective, et en particulier de l'émotion induite par les vidéos, nous avons élaboré une base de données de vidéos annotées selon l'émotion perçue par les visionneurs. Cette base, **LIRIS-ACCEDE**[3], contient un grand nombre de vidéos variées sous licence "Creative Commons" et pouvant donc être librement diffusées. Elle est constituée de deux types d'annotation : 10900 extraits vidéos d'une dizaine de secondes sont annotés globalement selon la valence (de l'émotion la plus négative à la plus positive) et l'arousal (de l'émotion la plus calme à la plus dynamique), et 66 films (36 heures) sont annotés de manière continue (chaque seconde) selon la valence et l'arousal, ainsi que la peur. La qualité de cette base a été reconnue d'une part par plusieurs publications dans les conférences et journaux du domaine de l'informatique affective [2,5,9,12,14,15,16], et d'autre part par son adoption comme données pour les tâches «Affective Impact of Movies» à MediaEval 2015, et «Emotional Impact of Movies» à MediaEval 2016, 2018 et 2019. Le nombre de téléchargements est actuellement de 439 (janvier 2020).

Dans le domaine de la robotique, nous avons par ailleurs proposé la base **Jacquard**[4] pour la détection dans des images de prises d'objets par des bras robotiques. Cette base est constituée de 54 485 scènes différentes à partir de 11 619 objets distincts avec un total de 4 967 454 annotations de prises. Elle a actuellement été téléchargée par 53 équipes (janvier 2020).

3. http ://liris-accede.ec-lyon.fr/
4. https ://jacquard.liris.cnrs.fr/