Contributions à la compréhension automatique de données visuelles

Emmanuel Dellandréa

Habilitation à Diriger des Recherches - Ecole Centrale de Lyon, LIRIS

Le 12 juin 2020

Incrustation vidéo

Membres du jury

Rapporteurs

Christine Fernandez-Maloigne, Prof., Université de Poitiers
Su Ruan, Prof., Université de Rouen
Benoit Huet, MCF HDR, EURECOM, Median Technologies

Examinateurs

Martha Larson, Prof., Université de Radboud (Pays-Bas)
Jenny Benois-Pineau, Prof., Université de Bordeaux
Georges Quénot, Dir. de recherche, LIG
Mohand Saïd Hacid, Prof., Université Lyon 1
Liming Chen, Prof., Ecole Centrale de Lyon

Incrustation vidéo

Plan de la présentation

- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives

Incrustation vidéo

Parcours professionnel

Diplôme d'ingénieur en Informatique (Ecole Polytechnique de l'Université de Tours)
DEA Signaux et Images en Biologie et Médecine (Université de Tours)

Doctorat en Informatique de l'Université de Tours

Analyse de signaux vidéos et sonores : application à l'étude de signaux médicaux.
(Encadrement : Pr. Nicole Vincent et Dr. Pascal Makris)

2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020...

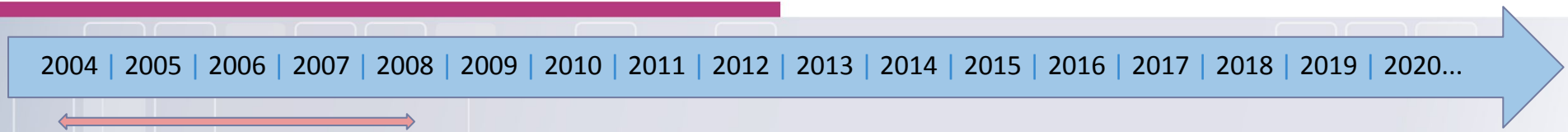
Maître de Conférences à l'Ecole Centrale de Lyon, laboratoire LIRIS

Attaché temporaire d'Enseignement et de Recherche (Université de Tours)

Allocataire de recherche et moniteur (Université de Tours)

Incrustation vidéo

Co-encadrement doctoral (thèses soutenues)



Zongzhe Xiao : *Recognition of emotions in audio signals*

Huanzhang Fu : *Contributions to generic visual object categorization*

Xi Zhao : *3D face analysis : landmarking, expression recognition and beyond*

Ningning Liu : *Contributions to generic and affective visual concept recognition*

Boyang Gao : *Contributions to music semantic analysis and its acceleration techniques*

Yoann Baveye : *Automatic prediction of emotions induced by movies*

Yuxing Tang : *Weakly Supervised Learning of Deformable Part Models and CNN for Object Detection*

Matthieu Grard : *Generic Instance Segmentation for Object-Oriented Bin-Picking*

Incrustation vidéo

Co-encadrement doctoral (thèses en cours)

- **Amaury Depierre (depuis 2017) :**

  ➜ Apprentissage profond pour une saisie adaptative par des bras robotiques

  financement CIFRE avec Siléane

- **Thomas Duboudin (depuis 2019) :**

  ➜ Apprentissage profond de simulation augmentée en vue d'applications aéroportées et terrestres

  financement CIFRE avec Thalès

Incrustation vidéo

Projets de recherche

2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020...

ACI Musicdiscover (Ircam, LTCI)

CNRS PICS MusicFinder (Département de Génie Electrique de l'Université Tsinghua)

ANR Omnia (LIG, Xerox XRCE)

ANR Videosense (EURECOM, LIG, LIF et GHANNI)

CHIST ERA Visen (Université de Surrey, Université de Sheffield et IRI)

FUI Pikaflex (Renault, Siléane)

Labcom Arès (Siléane) ...

CHIST ERA Learn-Real (Idiap et IIT)

Incrustation vidéo

Principales responsabilités

- **Directeur de l'Unité d'Enseignement de l'Informatique à l'Ecole Centrale de Lyon depuis 2010**
- **Responsable de l'équipe d'enseignement de l'Informatique depuis 2018**
- **Responsable du laboratoire LIRIS à l'ECL depuis janvier 2020**

Incrustation vidéo

Principales activités d'enseignement

- **Enseignements dans les 3 années du cursus d'ingénieur de l'ECL**
- **Matières fondamentales :**

  Algorithmique, programmation orientée objet, programmation web, ...
- **Matières liées aux activités de recherche :**

  Analyse de données, indexation vidéo, analyse multimédia, machine learning, deep learning, ...

Incrustation vidéo

Pour résumer

- MCF depuis 16 ans
- Co-encadrement de 10 thèses
- Participation au montage et réalisation de 8 projets collaboratifs
- Organisation de plusieurs événements scientifiques (dont “Emotional Impact of Movies” à MediaEval)
- Activité de relecture régulière (RI et CI)
- Publication de 16 articles de RI et une 30aine de CI
- Bénéficiaire de la PEDR (puis PES) depuis 2013

Incrustation vidéo

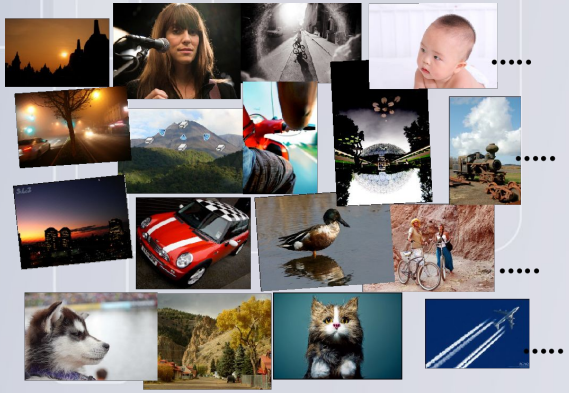
Plan de la présentation

- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives

Incrustation vidéo

Contexte

Croissance exponentielle de la quantité de données visuelles



Grande quantité de données en et hors ligne

100 heures de vidéos / min

500 millions de messages / jour

350 millions de photos / jour

➜Nécessité d'outils efficaces pour l'organisation, la recherche, la classification et l'interprétation

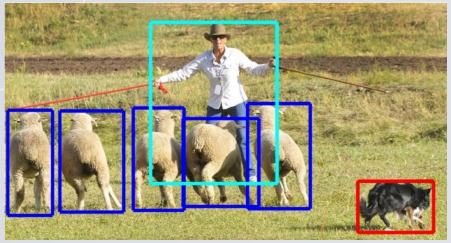
➜Emulation extrêmement importante dans les communautés de vision par ordinateur et d'apprentissage automatique

Incrustation vidéo

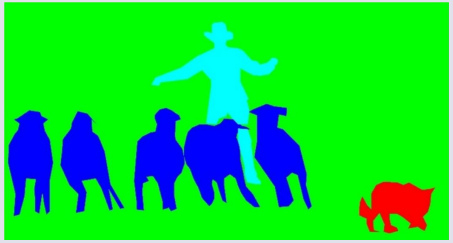
Contexte

Objectif : extraire automatiquement de l'information sémantique sur le contenu des images directement à partir des valeurs des pixels

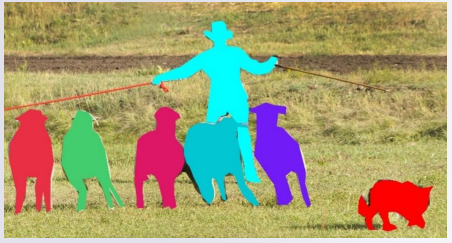
Classification d'images



Détection d'objets



Segmentation sémantique



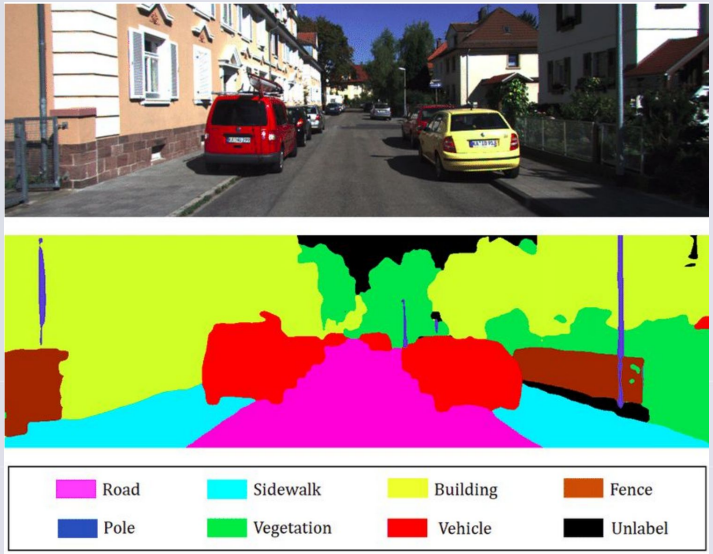
Détection d'instances

Mais aussi : génération automatique de légendes, détection de prises pour des bras robotiques, ...

Incrustation vidéo

Contexte

De nombreuses applications

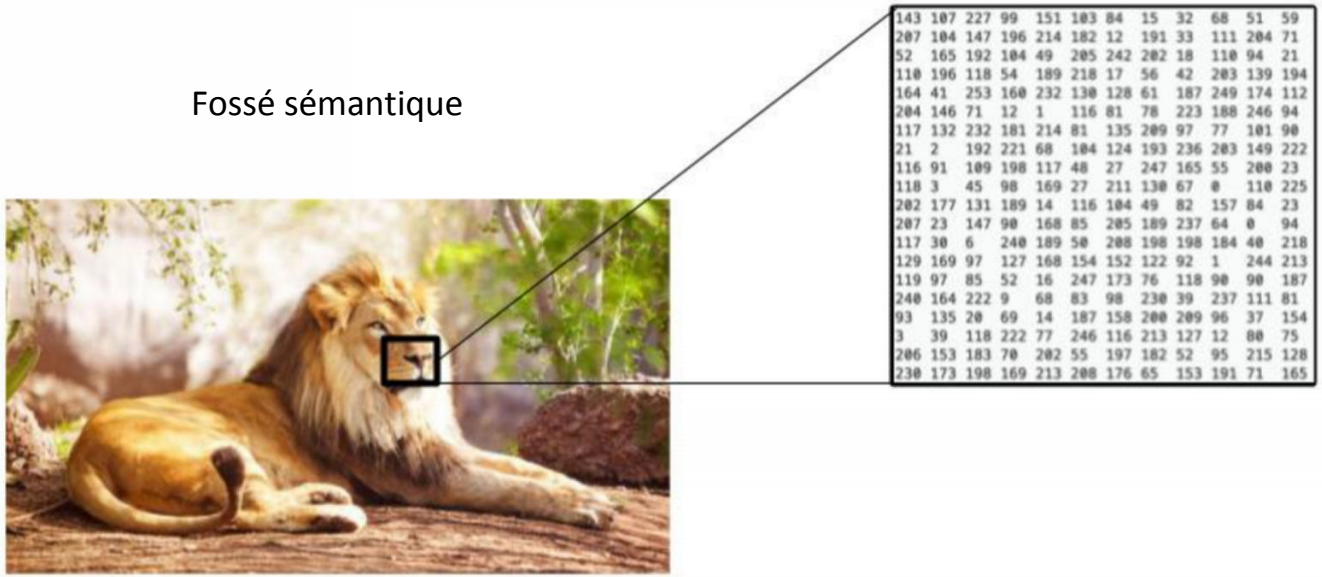


Incrustation vidéo

Contexte

Défis scientifiques

Fossé sémantique



Incrustation vidéo

Contexte

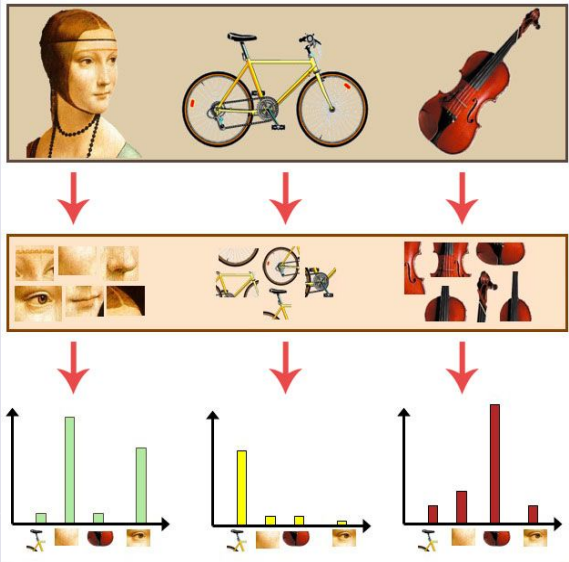
Défis scientifiques

Incrustation vidéo

Evolutions méthodologiques

Approche dominante dans les années 2000

1. Descripteurs locaux (HOG [DT05] et SIFT [Low04])
2. Agrégation de ces descripteurs par un apprentissage non supervisé (descripteurs globaux) (mots visuels [CDF + 04] ou vecteurs Fisher [PD07])
3. Classification par apprentissage supervisé (méthodes à noyaux SVM [CV95])

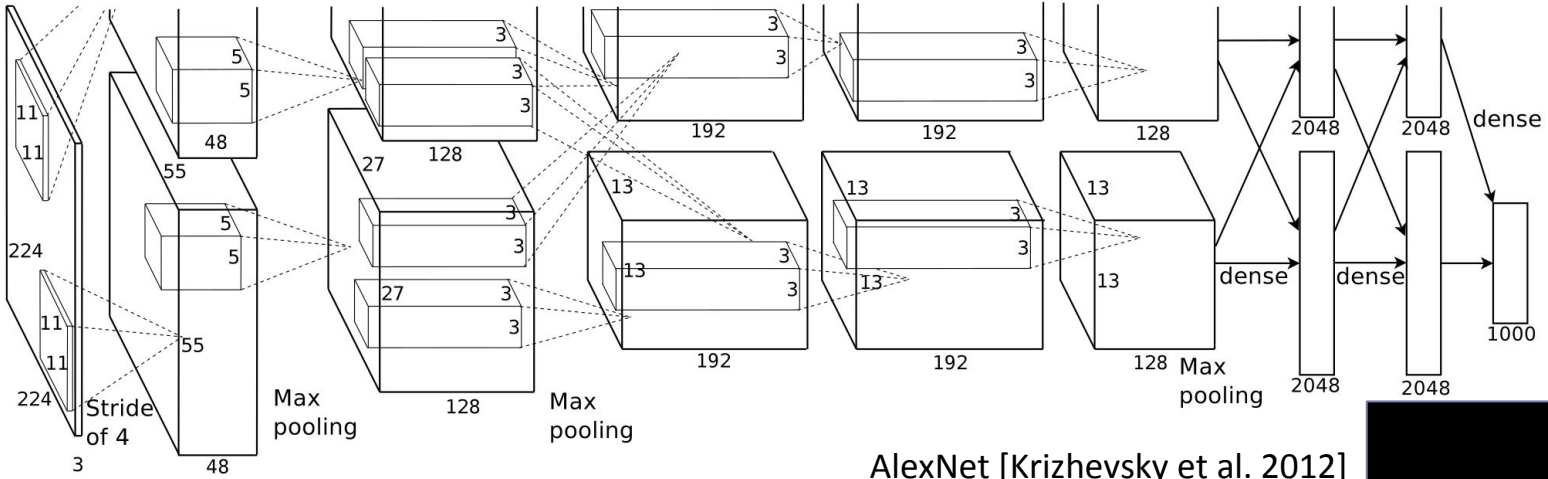


Incrustation vidéo

Evolutions méthodologiques

A partir de 2012 : avènement de l'apprentissage profond

➜ Succès exceptionnel d'un réseau convolutif à ILSVRC12

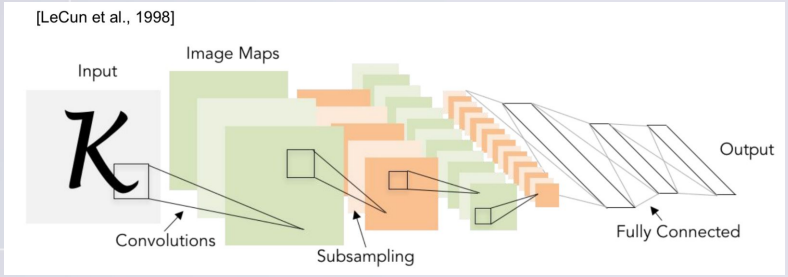


AlexNet [Krizhevsky et al. 2012]

Incrustation vidéo

Evolutions méthodologiques

Idée remontant aux années 90



Force des réseaux convolutifs :

➔ Apprentissage cohérent basé sur les données

- de représentations des images de bas et haut niveau sémantique (par les couches de convolution du réseau)
- du classifieur (par les couches entièrement connectées)

Incrustation vidéo

Evolutions méthodologiques

Essor de l'apprentissage profond possible par la convergence :

- de la puissance de calculs parallèles offerte par les cartes graphiques
- des quantités immenses de données disponibles (ImageNet)

Ouvre de nombreux nouveaux champs d'investigation :

- Comment apprendre de nouvelles tâches avec peu de données ?
- Interprétabilité des prédictions ?
- …

Incrustation vidéo

Compréhension automatique de données visuelles : Nos contributions

- **Classification d'images**
- **Détection d'objets dans les images**
- **Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos**
- **Analyse visuelle pour la robotique**

Incrustation vidéo

Plan de la présentation

- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives

Incrustation vidéo

Classification d'images

Objectif : associer automatiquement à l'image des étiquettes indiquant les concepts de haut-niveau sémantique présents

- des scènes (intérieur, extérieur, paysage, ...)
- des objets (voiture, animal, personne, ...)
- des événements (voyage, travail, ...)
- des émotions (joie, mélancolie, ...)

Nos contributions :

- Descripteur textuel pour la caractérisation des images
- Fusion multimodale

Incrustation vidéo

Descripteur textuel pour la caractérisation des images

Objectif : compléter les informations portées par les descripteurs visuels

{0A432C9F-1732-45E6-90F7-A6A7B75FA889}.jpg
Flickr user tags: peacock, bird, beautiful, pretty, feathers, waimea, waimeafalls, explore, animal, interestingness

Limites des principales approches existantes ➜ soit uniquement statistiques, soit tentent d'intégrer les relations sémantiques

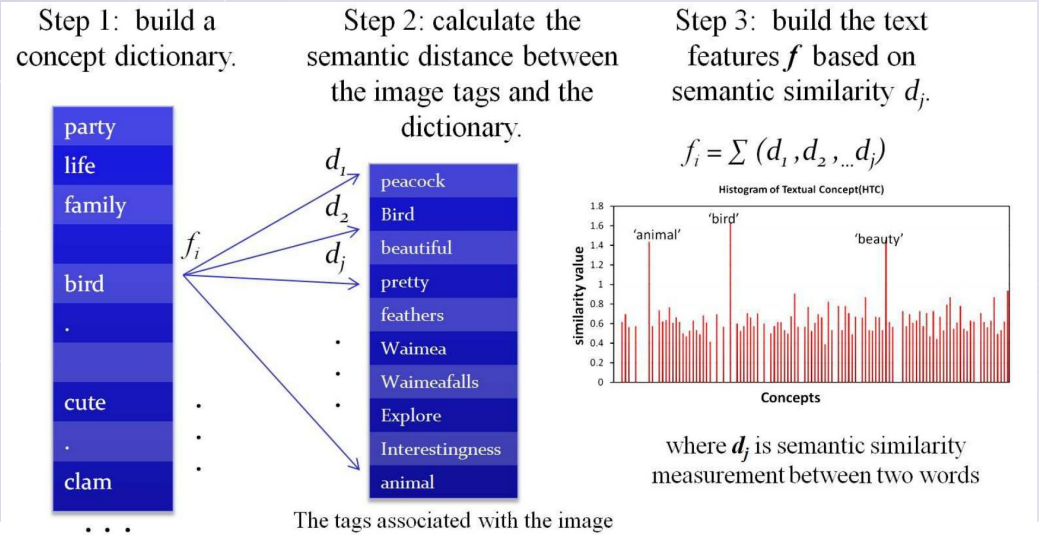
entre mots mais uniquement au sein du même document

Incrustation vidéo

Descripteur textuel pour la caractérisation des images

Notre proposition [Thèse de Ningning Liu] : Histograms of Textual Concepts (HTC)

Objectif : capturer les relations sémantiques entre concepts



Incrustation vidéo

Descripteur textuel pour la caractérisation des images

Expérimentations : tâche "Photo annotation" d'ImageCLEF en 2012

Objectif : annoter automatiquement 10 000 images selon 94 concepts visuels (15 000 images d'entraînement)

Soumissions :

- Participation de 18 équipes internationales
- 80 soumissions dont 17 exclusivement textuelles

Incrustation vidéo

Descripteur textuel pour la caractérisation des images

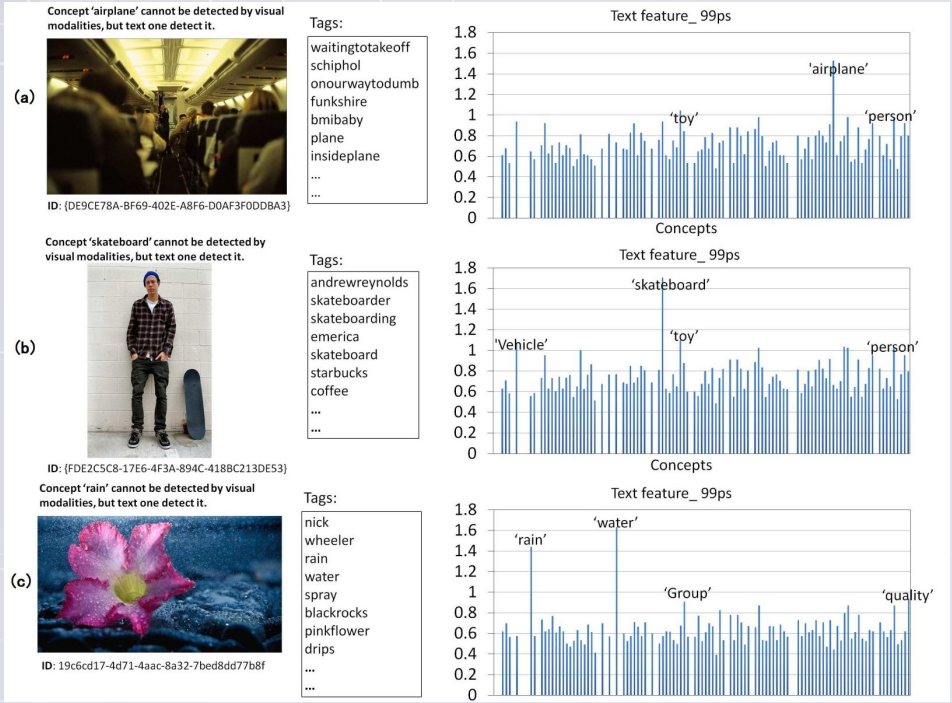
Résultats

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		

Incrustation vidéo

Descripteur textuel pour la caractérisation des images

Résultats



Publié dans : ACII 2011, CVIU (2013), Plos One (2017)

Incrustation vidéo

Fusion multimodale

Objectif : combiner plusieurs sources d'information afin d'en tirer le meilleur parti pour les concepts visuels à reconnaître

Principales stratégies de fusion :

- Au niveau des descripteurs (fusion précoce) [PBAC + 17]
- Au niveau du score des classifieurs (fusion tardive) [SWS05, TSBB + 14]
- A des niveaux intermédiaires, par exemple au niveau de noyaux [AQG07]

Incrustation vidéo

Fusion multimodale

Notre proposition [Thèse de Ningning Liu] : "Selective Weighted Late Fusion" (SWLF)

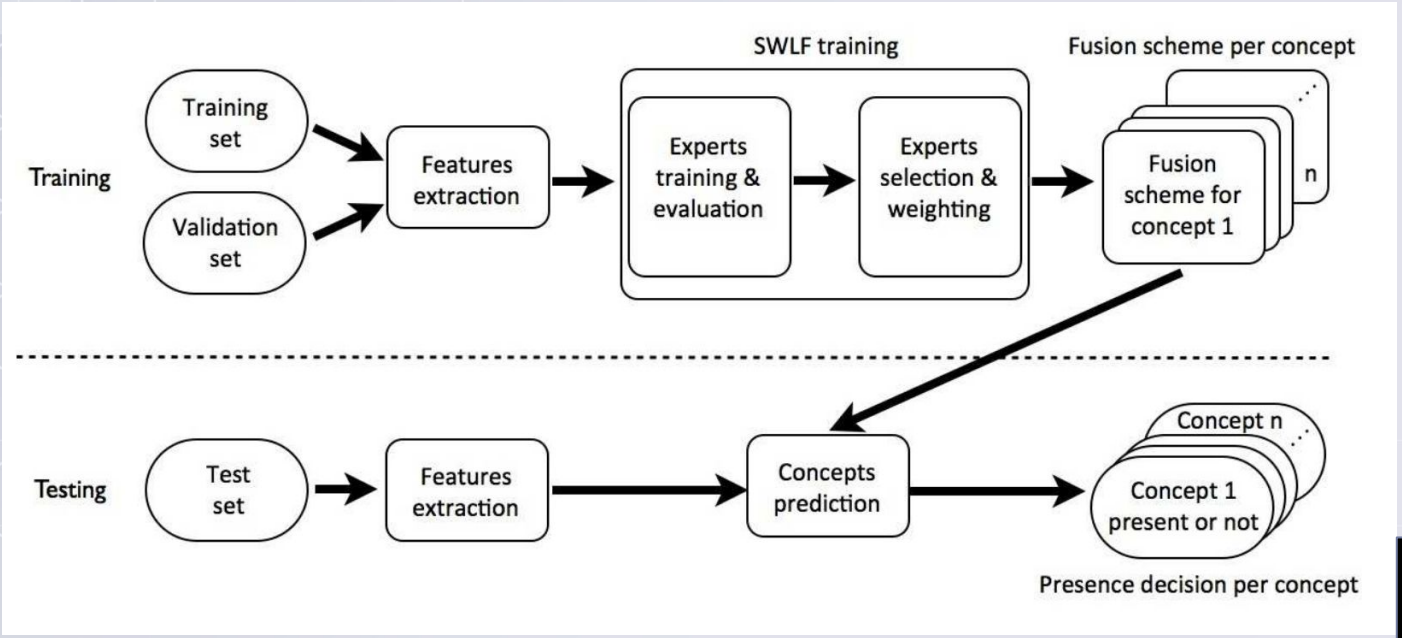
Hypothèses :

- ➔ Le score de classification à partir d'un type de descripteur (classifieur expert) doit être pondéré en fonction de sa qualité intrinsèque
- ➔ Différents concepts visuels peuvent nécessiter différents types de descripteurs pour permettre leur reconnaissance de manière efficace

Incrustation vidéo

Fusion multimodale

SWLF :



Incrustation vidéo

Fusion multimodale

Expérimentations : tâche "Photo annotation" d'ImageCLEF en 2012

- 11 descripteurs textuels issus de HTC
- 24 descripteurs visuels locaux et globaux
- Classifieurs SVM

Incrustation vidéo

Fusion multimodale

Résultats

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

Publié dans : ECCV 2012 Workshop, CVIU (2013), Springer (2014)

Incrustation vidéo

Plan de la présentation

- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives

Incrustation vidéo

Détection d'objets dans les images

Objectif : localiser les objets dans les images (boîtes englobantes)

Principale difficulté :

➔ Disposer d'un nombre suffisant d'images annotées manuellement avec des boîtes englobantes précisant la nature et la localisation des objets

Nos contributions :

- Détection faiblement supervisée (annotations au niveau global des images)
- Détection semi-supervisée (seule une partie des catégories à reconnaître possède des annotations au niveau des objets)

Incrustation vidéo

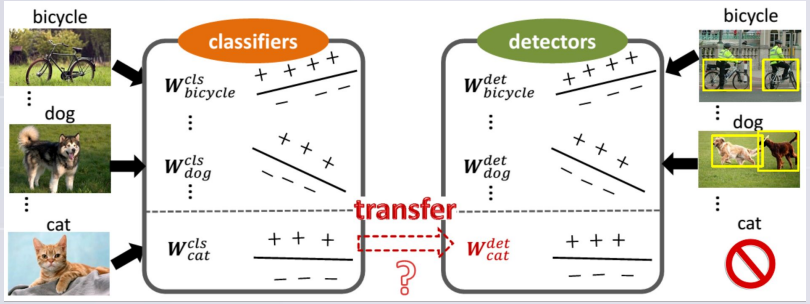
Détection d'objets semi-supervisée

Positionnement du problème : 2 types de catégories

- Catégories «entièrement annotées» : annotation des labels au niveau global de l'image et au niveau des boîtes englobantes
- Catégories "faiblement annotées" : labels uniquement au niveau global

Objectif :

➔ Convertir des classifieurs en détecteurs pour les catégories faiblement annotées



Incrustation vidéo

Détection d'objets semi-supervisée

Large Scale Detection through Adaptation (LSDA) [HGT + 14]

➔ Transférer les connaissances de différences entre les classifieurs et détecteurs pour des catégories fortes aux classifieurs pour des catégories faibles et proches visuellement

Mesure de similarité entre catégories : distance dans l'espace des poids de la couche fc8 du réseau CNN (AlexNet)

Principe :

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

$\overrightarrow{w_j^d}$ poids de fc8A du détecteur pour la catégorie *j*

$\Delta_{B_i^j}$ différences de poids de la couche fc8 de la *i*ème catégorie voisine dans l'ensemble B pour la catégorie *j*

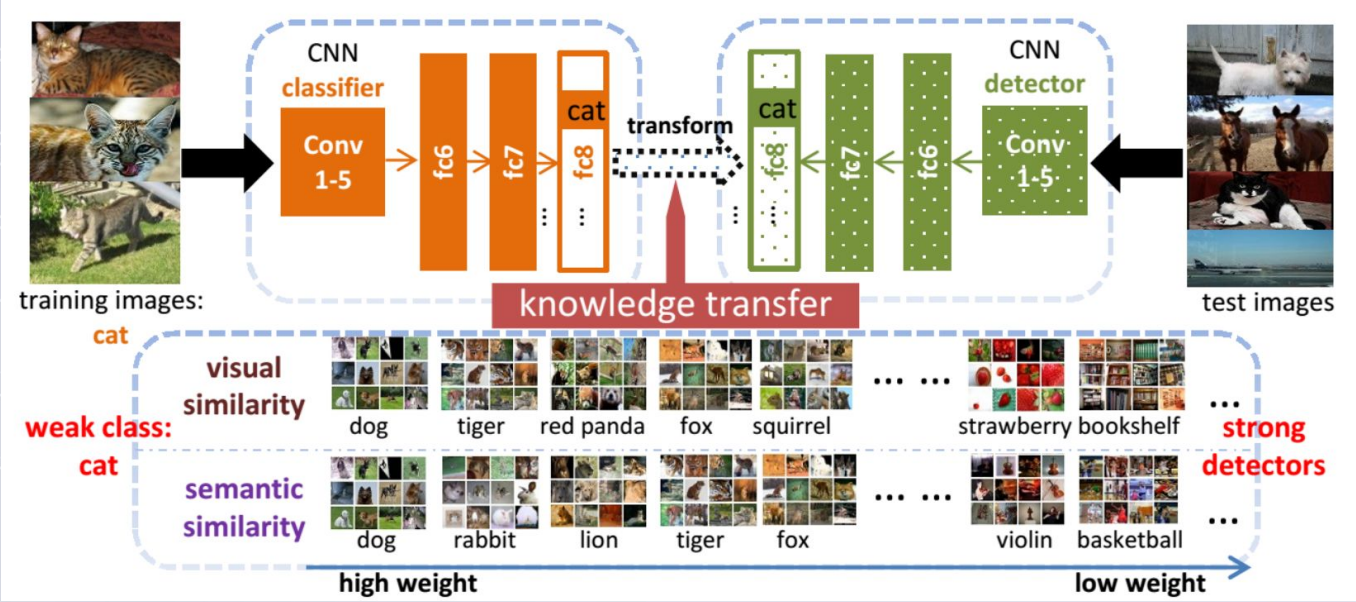
A : catégories faibles, B catégories fortes

Incrustation vidéo

Détection d'objets semi-supervisée

Nos contributions [Thèse de Yuxing Tang] :

→ Similarité visuelle plus adaptée + ajout d'une similarité sémantique



Incrustation vidéo

Détection d'objets semi-supervisée

Transfert de connaissance par similarité visuelle :

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

Transformation des poids :

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

Incrustation vidéo

Détection d'objets semi-supervisée

Transfert de connaissance par similarité similarité sémantique :

- ➔ Représentation des catégories par plongement lexical (word2vec)
  (1 catégorie ⇒ 1 vecteur de dimension 300)
- ➔ Vecteur égal à la somme normalisée de chaque terme du synset Wordnet
- ➔ Distance sémantique entre chaque catégorie $j \in A$ et $i \in B$: norme *l2*
  ds(j, i) de chaque paire
- ➔ Similarité sémantique : inverse de la distance sémantique

Transformation des poids :

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

Incrustation vidéo

Détection d'objets semi-supervisée

Modèle complet de transfert de connaissance :

- Similarité visuelle au niveau global de l'image
- Similarité sémantique au niveau des objets

→ Combinaison des deux similarités complémentaires :

$$s = intersect[\alpha s_v + (1 - \alpha) s_s]$$

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.

Incrustation vidéo

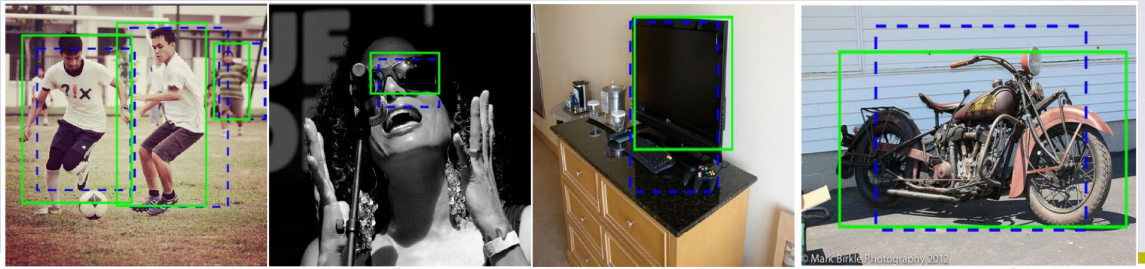
Détection d’objets semi-supervisée

Propositions de boîtes englobantes par Selective Search

Problème : prédiction de position et taille approximative

➜ Solution : Transfert de connaissance pour la régression des boîtes englobantes

- Apprentissage de la régression pour les catégories fortes
- Transfert de ces connaissances aux catégories faibles en utilisant les mesures de similarité

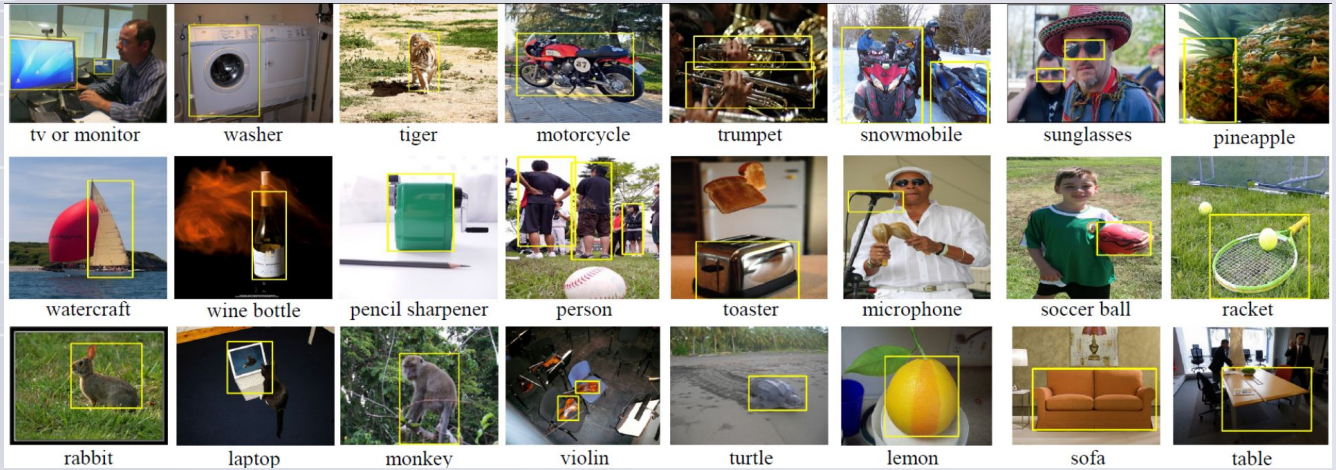


Incrustation vidéo

Détection d’objets semi-supervisée

Expérimentations

- Jeu de données ILSVRC2013 couvrant 200 catégories d’objets
  - Annotations au niveau global des images pour les 200 catégories
  - Annotation des boîtes englobantes pour les 100 catégories



Incrustation vidéo

Détection d'objets semi-supervisée

Résultats

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. - ≤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

Publié dans : CVPR 2016, IEEE PAMI (2018)

Incrustation vidéo

Plan de la présentation

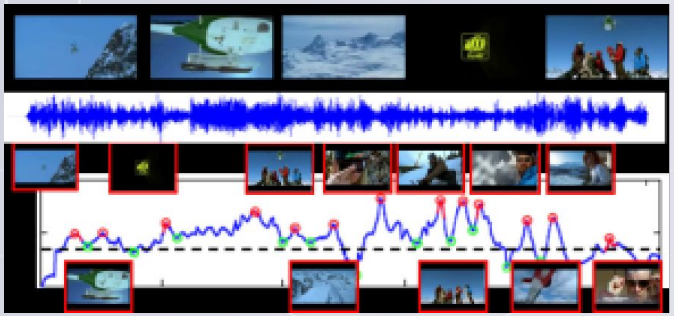
- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives

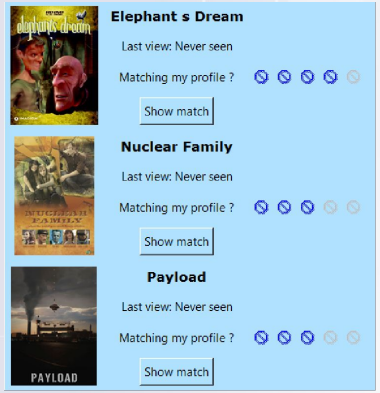
Incrustation vidéo

Prédiction de l’impact émotionnel des vidéos

Prédire l’émotion suscitée par un film utile pour :

- ➔ La recommandation automatique basée sur l'émotion
- ➔ L'aide à la création de contenus audiovisuels
- ➔ Le filtrage de vidéos au contenu pouvant être inapproprié





Incrustation vidéo

Prédiction de l'impact émotionnel des vidéos

- **Importants progrès réalisés en vision par ordinateur et apprentissage automatique notamment pour la compréhension de scènes visuelles**
- **Etape suivante :** modéliser et reconnaître les concepts affectifs

But : doter les ordinateurs de capacités de perception semblables à celles des humains

➜ Challenge très relevé, notamment en raison de la complexité et de la nature subjective des émotions

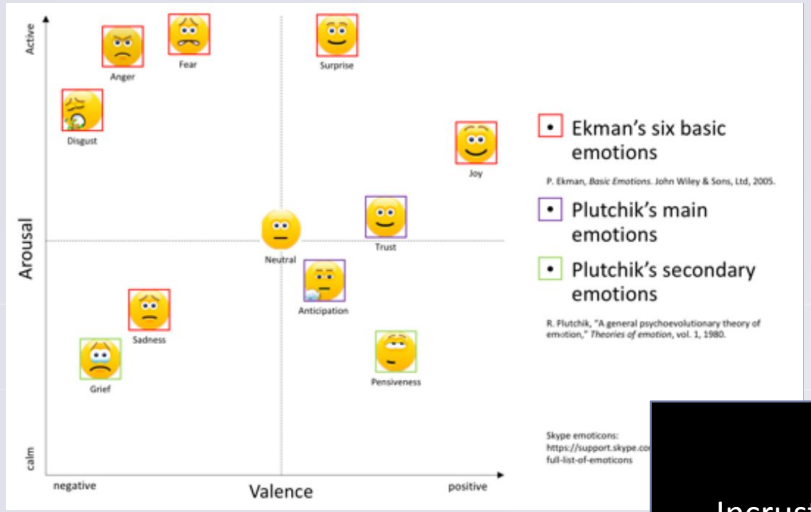
Incrustation vidéo

Prédiction de l’impact émotionnel des vidéos

Nos contributions [Thèse de Yoann Baveye] :

- ➔ Jeu de donnés LIRIS-ACCEDE
- ➔ Modèle spatio-temporel profond pour la prédiction de l’impact émotionnel des vidéos

LIRIS-ACCEDE (liris-accede.ec-lyon.fr)



Incrustation vidéo

LIRIS-ACCEDE

Motivations : Lever les limites des jeux de données existants

- Petite taille (quelques centaines d'éléments)
- Faible diversité de contenu
- Problèmes de droits d'auteur

➜ **Chaque chercheur constitue son propre jeu de données**

➜ **Pas de possibilité d'évaluer les méthodes les unes par rapport aux autres**

Incrustation vidéo

LIRIS-ACCEDE

Grand nombre de vidéos variées sous licence Creative Commons

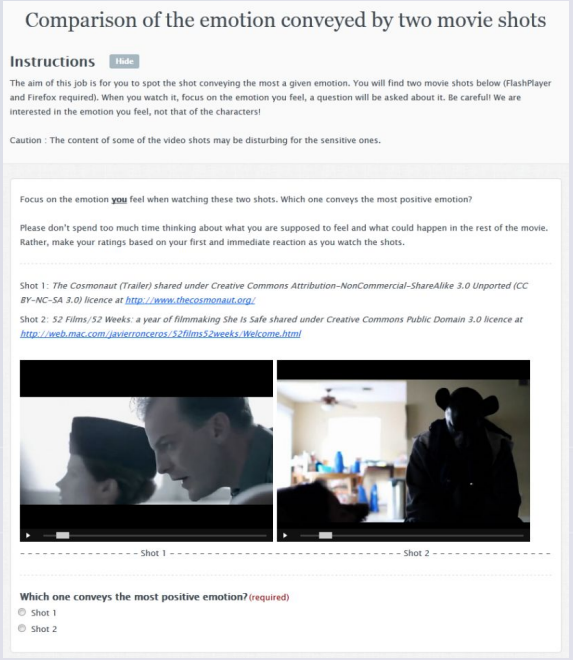
- **2 types d'annotations :**
  - 10900 extraits vidéos d'une dizaine de secondes annotés globalement selon la valence et l'arousal
  - 66 films (36 heures) annotés de manière continue (chaque seconde) selon la valence et l'arousal, ainsi que la peur
- **Organisée en 6 collections dont**
  - 1 utilisée comme jeu de données pour la tâche "Affective impact of movies" à MediaEval 2015
  - 3 utilisées pour la tâche "Emotional impact of movies" à MediaEval 2016, 2017 et 2018

Incrustation vidéo

LIRIS-ACCEDE

Collecte de la vérité terrain pourles annotations globales

- ➔ Crowdsourcing pour annoter des paires de vidéos (par comparaison)

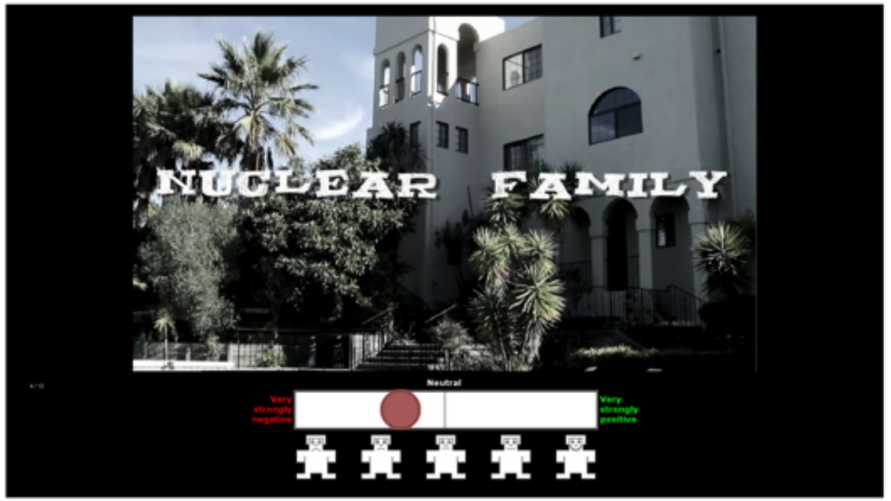


Incrustation vidéo

LIRIS-ACCEDE

Collecte de la vérité terrain pour les annotations continues

➔ Environnement contrôlé avec interface de visualisation et joystick pour indiquer le niveau de valence /arousal

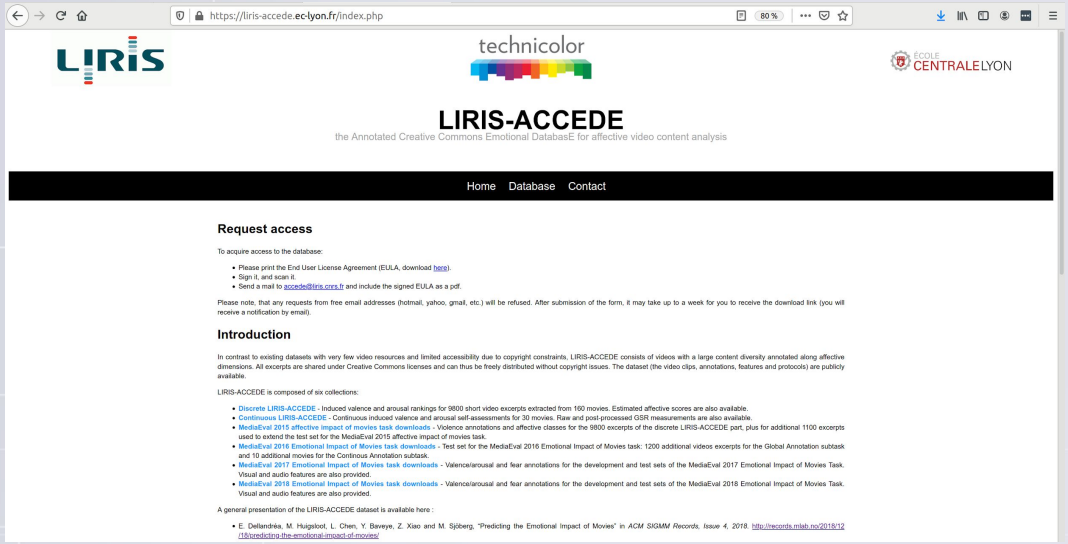


Screenshot during the annotation along the valence axis

Incrustation vidéo

LIRIS-ACCEDE

Environ 480 téléchargements (juin 2020)



Publié dans : ACII 2013 et 2015, IEEE Trans. Affective Computing (2015), ACM SIGMM Records (2018)

Incrustation vidéo

Modèle de prédiction spatio-temporel

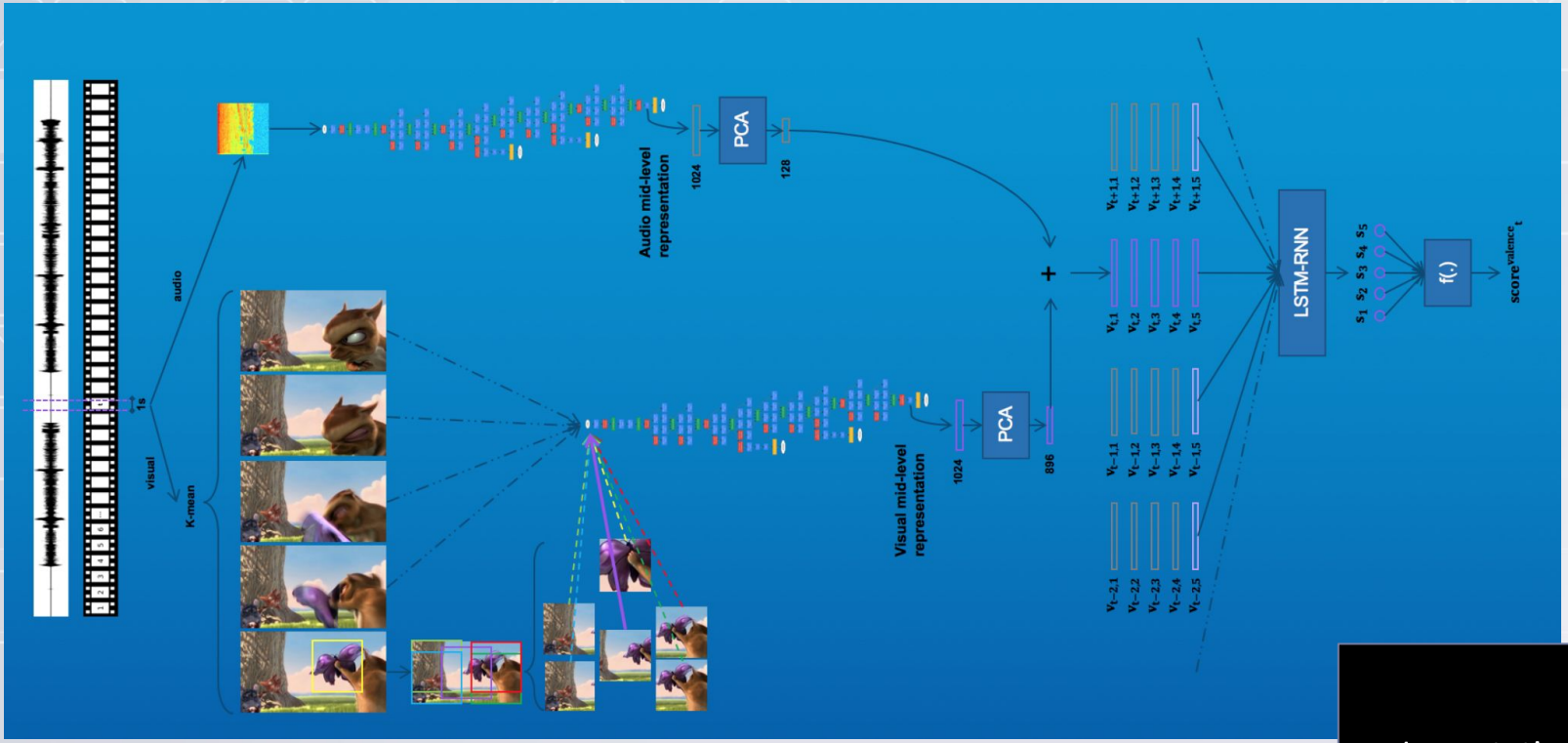
Objectif du modèle :

- Tirer partie des très bonnes performances des réseaux convolutifs pour la classification d'images / détection d'objets
- Intégrer la modalité audio
- Intégrer la temporalité (LSTM)

  ➔ 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

Incrustation vidéo

Modèle de prédiction spatio-temporel



Incrustation vidéo

Modèle de prédiction spatio-temporel

Expérimentations :

➔ 30 vidéos de la collection Continuous LIRIS-ACCEDE

- 23 pour l'apprentissage
- 7 pour le test

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**

Incrustation vidéo

Plan de la présentation

- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives

Incrustation vidéo

Analyse visuelle pour la robotique

Objectif : doter les robots

- d'une vision artificielle (observer et de comprendre la scène)
- d'une intelligence (acquérir de nouvelles capacités ou s'adapter aux changements d'environnements)

→ Application au Picking/Kitting sur des bases robotiques afin de les rendre flexibles, adaptables et autonomes

Nos contributions :

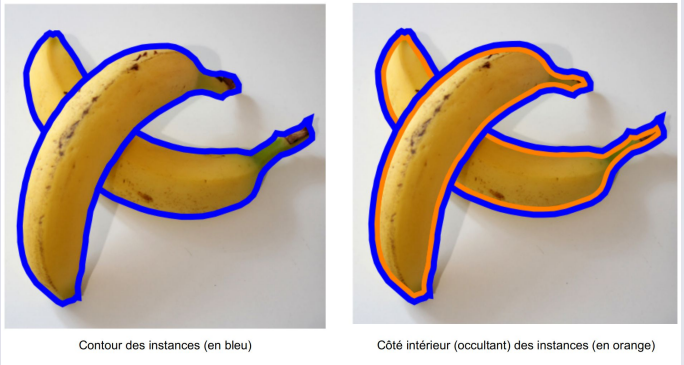
- Localisation d'instances d'objets dans un vrac homogène
- Prédiction de prises pour des bras robotiques

Incrustation vidéo

Localisation d’instances d’objets dans un vrac

Objectif : délimiter des instances d’objets et comprendre leur disposition spatiale à partir d’une unique image RGB sans modèle explicite des objets

➜ Facilite la prise par le robot



Contour des instances (en bleu)

Côté intérieur (occultant) des instances (en orange)

Nécessité de porter l’attention au niveau des pixels des instances (représentations dépendantes de la position)

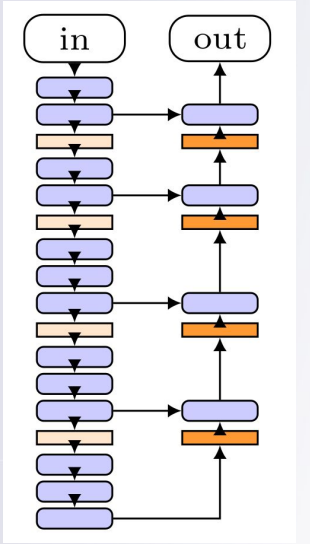
Problème : noaux convolutifs des CNN invariants à la translation

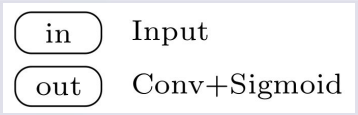
Incrustation vidéo

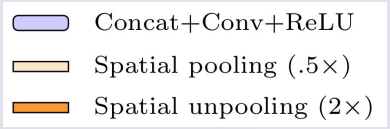
Localisation d'instances d'objets dans un vrac

➜ Utilisation d'un réseau encodeur-décodeur résiduel (RED) [CZP + 18]

- inférer les labels au niveau des pixels en combinant graduellement
  - des informations au niveau des objets à faible résolution
  - des indices locaux à résolution plus élevée
- le décodeur sur-échantillonne les représentations latentes de l'encodeur







Incrustation vidéo

Localisation d’instances d’objets dans un vrac

Innovations apportées au niveau de l’encodeur profond

➜ représentations dépendantes de la position à faible résolution des catégories d’objets variées

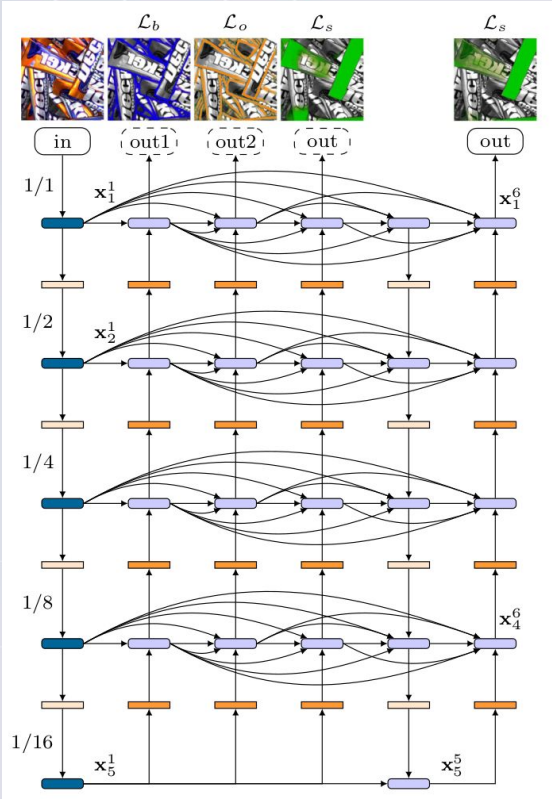
Problème : efficace pour des objets variés, mais pas pour des agencements homogènes denses

Incrustation vidéo

Localisation d'instances d'objets dans un vrac

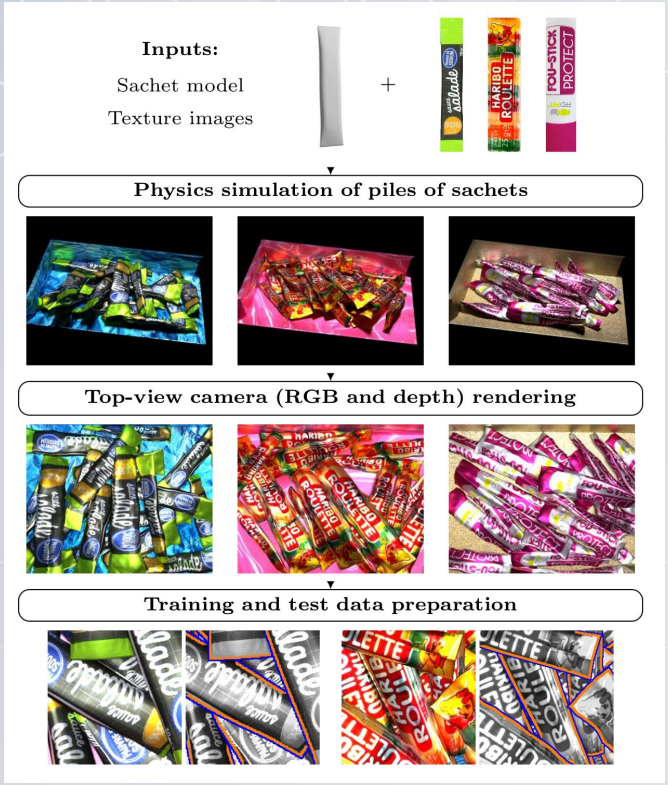
Proposition :

[Thèse de Matthieu Grard]



Incrustation vidéo

Localisation d'instances d'objets dans un vrac



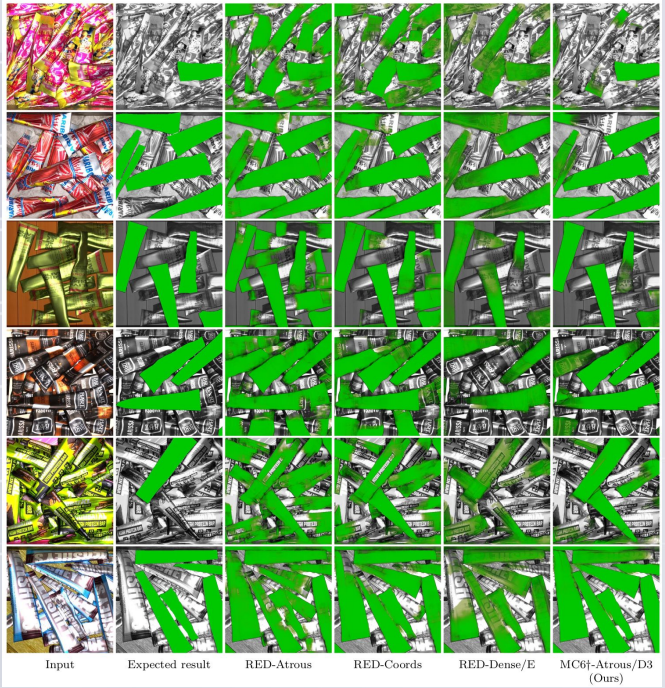
Jeu de données Mikado :

- 16960 images synthétiques réalistes de taille 640x512
- 507186 instances
- Moyenne de 30 instances par image

Incrustation vidéo

Localisation d'instances d'objets dans un vrac

Résultats :



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**

Publié dans IJCV (2020)

Incrustation vidéo

Prédiction de prises pour des bras robotiques

Objectif : prédiction des paramètres de prises (pince à mâchoires parallèles) sur les objets présents dans des images RGB

➔ **De nombreux domaines d'application :**

Industrie, logistique, interactions avec des humains, tri des déchets, ...

Difficulté :

➔ Pas de modèle explicite des objets

Incrustation vidéo

Prédiction de prises pour des bras robotiques

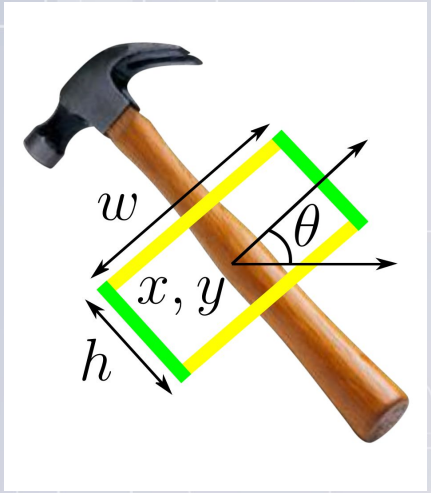
Solution de l'état de l'art [ZLZ + 18] :

- Boîtes d'ancrage dans les images avec orientations multiples
- Paramètres de prise prédits à partir de cette orientation de référence
- Réseau entièrement convolutif pour prédire par régression ces paramètres ainsi qu'un score de qualité de prise

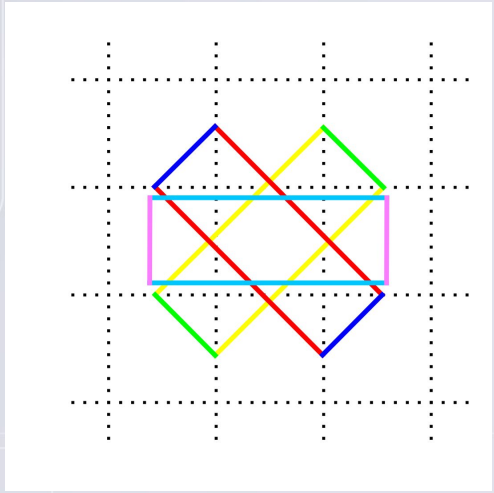
Incrustation vidéo

Prédiction de prises pour des bras robotiques

Représentation d'une prise :



Prise caractérisée par 5 valeurs



3 boîtes d'ancrage centrées sur le même pixel, orientées selon 3 directions

boîte d'encrage

$$g_a = (g_{ax}, g_{ay}, g_{aw}, g_{ah}, g_{a\theta})$$

déformation

$$\delta = (\delta_x, \delta_y, \delta_w, \delta_h, \delta_\theta)$$

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

Incrustation vidéo

Prédiction de prises pour des bras robotiques

Limitation : prédiction de la qualité de la prise uniquement dépendante de l'information dans l'image (non liée à la prédiction des valeurs de la prise)

Notre contribution [Thèse d'Amaury Depierre] : ajout d'une dépendance directe entre la prédiction des valeurs de la prise et le score caractérisant la qualité de la prise

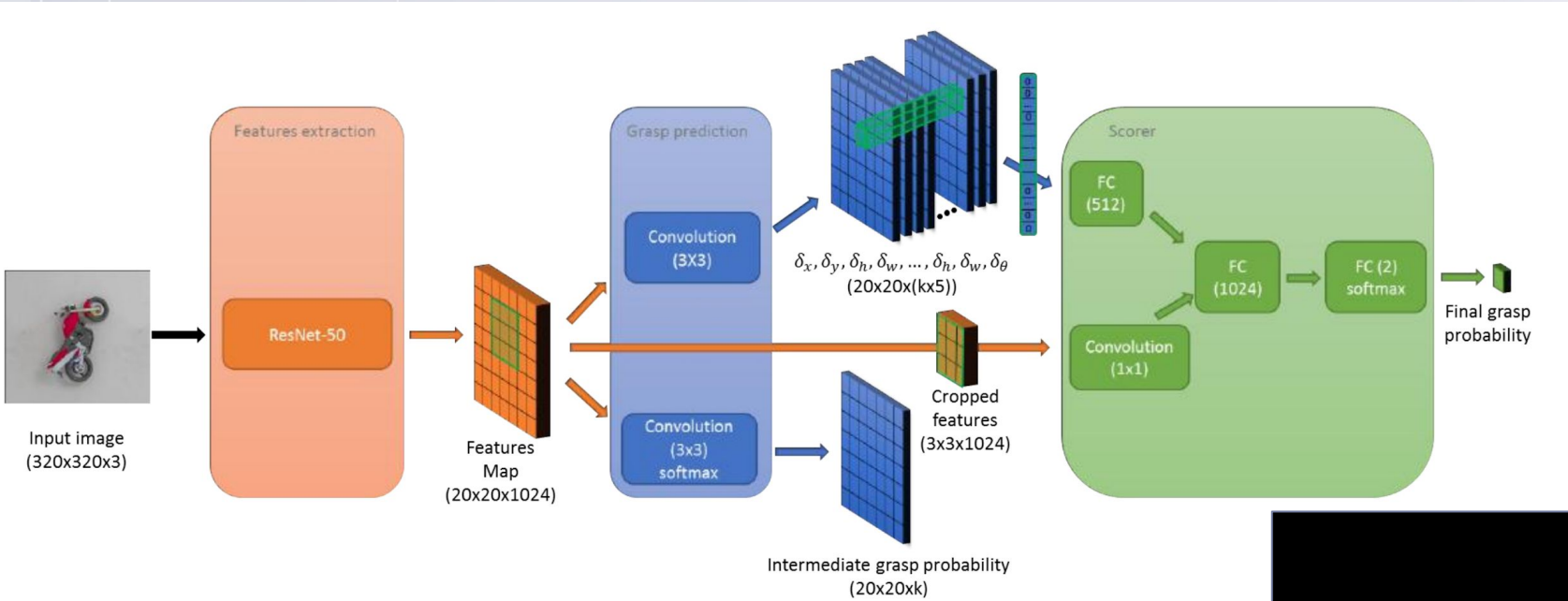
➔ **Réseau de score**

- Prédiction d'une probabilité décrivant la qualité de la prise proposée
- Utilisation de l'estimation du score pour améliorer la qualité de la régression

Incrustation vidéo

Prédiction de prises pour des bras robotiques

Notre approche :



Incrustation vidéo

Prédiction de prises pour des bras robotiques

Jeu de données de l'état de l'art : Cornell Grasping Dataset

1035 images de 280 objets différents, environ 10 000 annotations de prises

Limitations :

- Faible quantité de données limitant les performances de généralisation des modèles pré-entraînés
- Annotations humaines parfois imprécises et/ou incomplètes

Notre proposition : jeu de données **Jacquard**

- 54 485 scènes simulées réalistes à partir de 11 619 objets distincts
- 4 967 454 annotations de prises

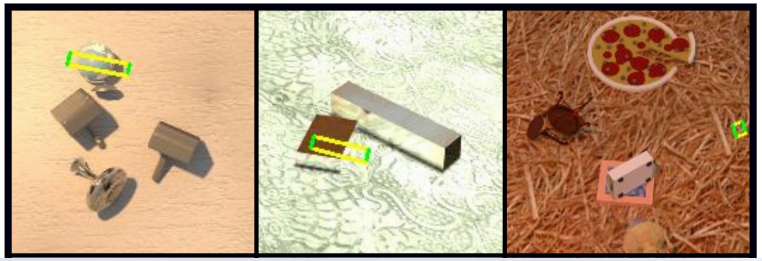
Incrustation vidéo

Prédiction de prises pour des bras robotiques

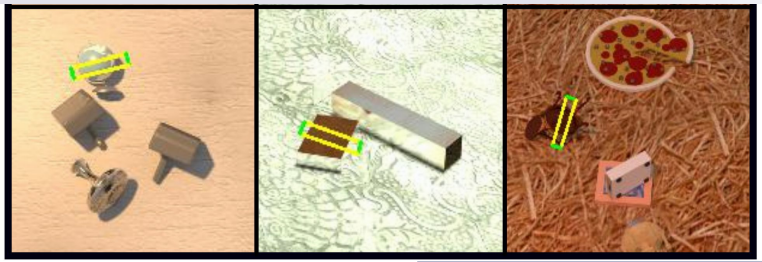
Résultats :

	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%**

Zhou et al.



Notre approche

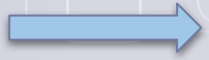


Critère de performance : Jaccard Matching (une prise est considérée bonne si elle est proche de la prise réelle avec des valeur seuil de 30°pour l'angle et 25% pour l'intersection sur l'union)

Incrustation vidéo

Plan de la présentation

- **Résumé du parcours professionnel**
- **Principales contributions scientifiques**
  - Contexte
  - Classification d'images
  - Détection d'objets dans les images
  - Analyse visuelle pour la prédiction de l'impact émotionnel des vidéos
  - Analyse visuelle pour la robotique
  - Conclusion et perspectives



Incrustation vidéo

Conclusion

Contributions pour la *Compréhension automatique de données visuelles* :

- **Classification d'images**
  ➜ Descripteurs parcimonieux, descripteurs textuels, fusion multimodale
- **Détection d'objets dans les images**
  ➜ Modèles de détection avec apprentissage faiblement et semi-supervisé
- **Prédiction de l'impact émotionnel de vidéos**
  ➜ Jeu de données LIRIS-ACCEDE et modèle spatio-temporel profond
- **Analyse visuelle pour la robotique**
  ➜ Localisation d'instances et prédiction de prises

Incrustation vidéo

Conclusion

➜ Travaux en collaboration avec *Liming Chen*, des doctorants et post-doctorants

- Classification d'images : ***Huanzhang Fu et Ningning Liu***
- Détection d'objets dans les images : ***Yuxing Tang***
- Prédiction de l'impact émotionnel de vidéos : ***Yoann Baveye***
- Analyse visuelle pour la robotique : ***Matthieu Grard et Amaury Depierre***

Mais également avec des partenaires académiques et industriels dans le cadre de projets nationaux et internationaux

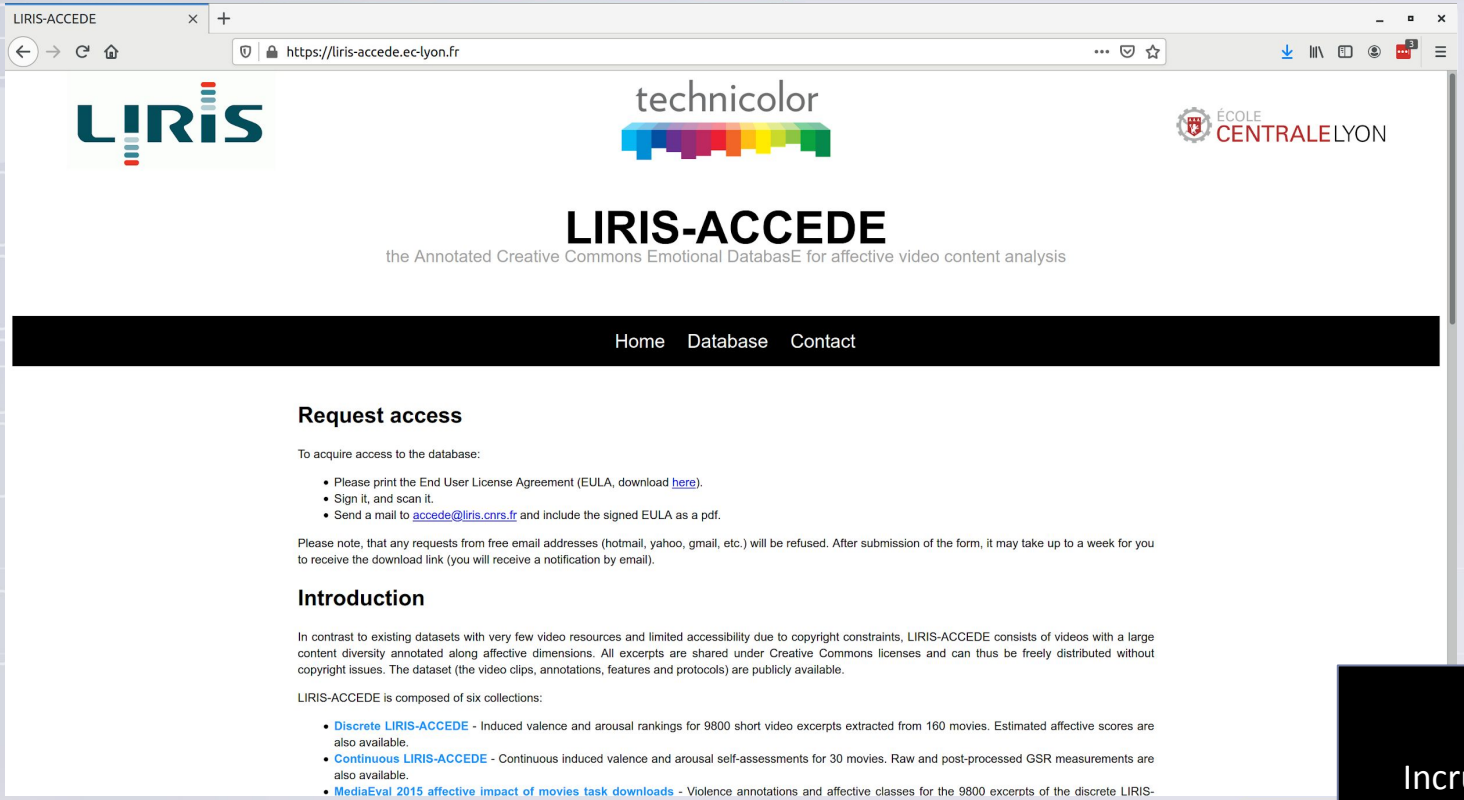
Incrustation vidéo

Dissémination

- **Contributions méthodologiques pour faire avancer l'état de l'art**
- **Contributions à la communauté en proposant de nouveaux jeux de données de grande dimension**
  - LIRIS-ACCEDE : émotions induites par les vidéos
  - Jacquard : détection de prises
  - Mikado : segmentation d'instances dans des vracs

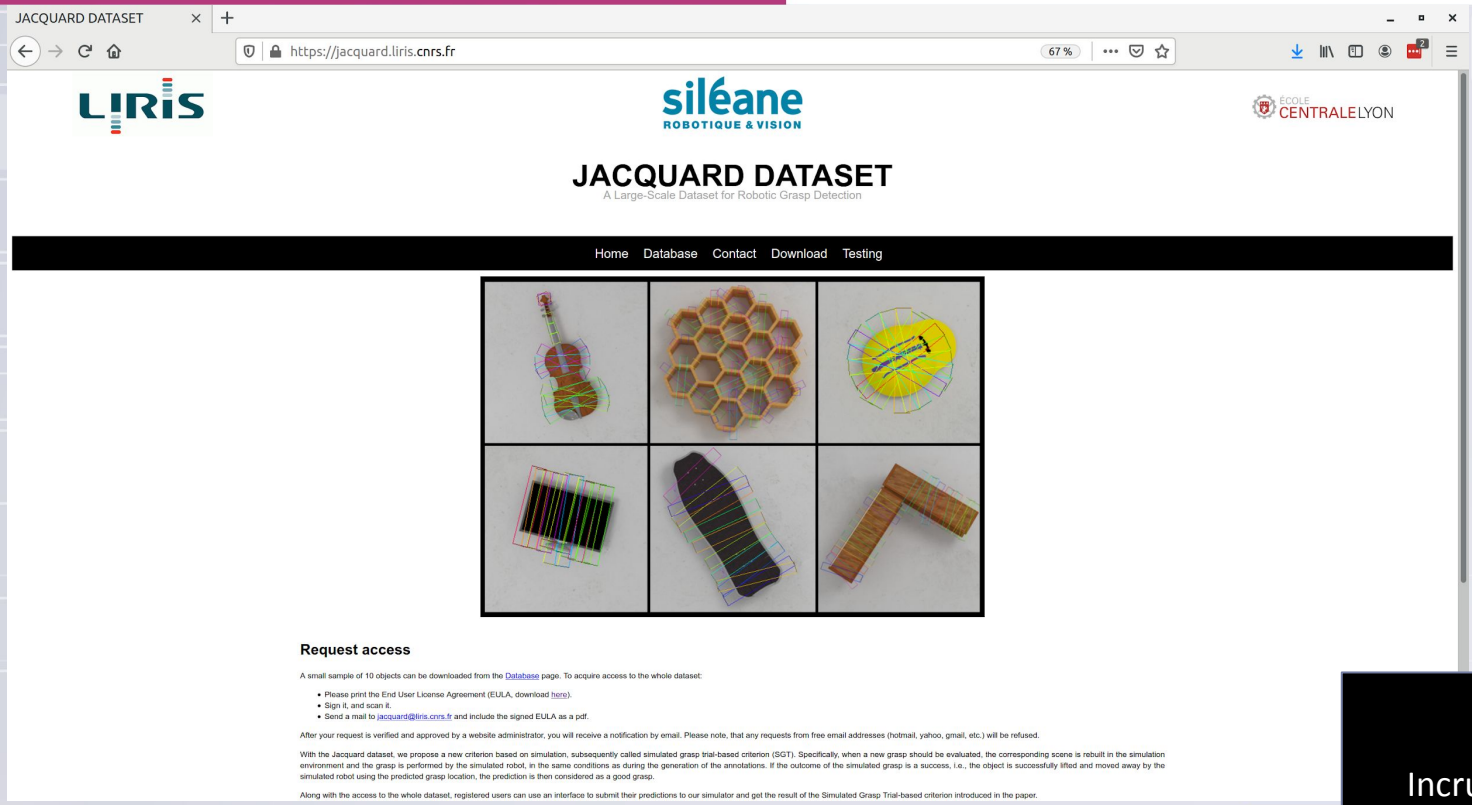
Incrustation vidéo

Site LIRIS-ACCEDE



Incrustation vidéo

Site Jacquard



Incrustation vidéo

Site Mikado

Mikado Dataset

https://mikado.liris.cnrs.fr 80 %

Mikado Dataset

A Synthetic Dataset of Dense Homogeneous Object Layouts
for Occlusion-aware Instance Segmentation

Home Dataset Download Contact

Collaborators

- Matthieu Grard, Siléane & Ecole Centrale de Lyon, LIRIS, France
- Emmanuel Dellandréa, Ecole Centrale de Lyon, LIRIS, France
- Liming Chen, Ecole Centrale de Lyon, LIRIS, France

Citation

If you use the Mikado dataset in your research, please cite the following paper:

- M. Grard, E. Dellandrea, and L. Chen, "Deep Multicameral Decoding for Localizing Unoccluded Object Instances from a Single RGB Image" in *International Journal of Computer Vision (IJCV)*, 2020. DOI: https://doi.org/10.1007/s11263-020-01323-0

This work was supported by

LIRIS siléane ÉCOLE CENTRALE LYON

Incrustation vidéo

Perspectives

Travaux futurs en continuité des activités actuelles

Problématique de la famine des données :

- GAN, Adaptation/généralisation de domaines (thèse de Thomas Duboudin)

  ➜ *Comment réduire le "reality gap" (détection d'objets dans des images aériennes) ?*

- Few-Shot Learning, Meta Learning (thèse d'Amaury Depierre, nouvelle thèse à venir)

  ➜ *Tâches de classification/régression peu similaires et données disponibles qu'au fur et à mesure au cours du temps*

Incrustation vidéo

Perspectives

Applications robotiques :

- Rendre possible la capitalisation de connaissances d'un robot et le transfert de connaissances entre robots pour permettre leurs permettre d'être flexibles, adaptables et autonomes dans des contextes instables et évolutifs (projet région FAIR Wastes)

- Saisie d'objets déformables (nouvelle thèse à venir, projet CHIST ERA Learn Real)

Incrustation vidéo

Merci de votre attention ! Questions ?

Contributions à la compréhension automatique de données visuelles

Emmanuel Dellandréa

Habilitation à diriger des recherches - Ecole Centrale de Lyon, LIRIS

Le 12 juin 2020

Rapporteurs

Christine Fernandez-Maloigne, Prof., Université de Poitiers
Su Ruan, Prof., Université de Rouen
Benoit Huet, MCF HDR, EURECOM, Median Technologies

Examinateurs

Martha Larson, Prof., Université de Radboud
Jenny Benois-Pineau, Prof., Université de Bordeaux
Georges Quénot, Dir. de recherche, LIG
Mohand Saïd Hacid, Prof., Université Lyon 1
Liming Chen, Prof., ECL

Incrustation vidéo

Matériels supplémentaires

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Représentation parcimonieuse :

- Représentation fidèle d'un signal considéré comme une combinaison linéaire d'atomes
- Ces atomes constituent un dictionnaire de dimension très supérieure à celle du signal lui-même

→ **Outil puissant pour acquérir, représenter et compresser des signaux de grande dimension**

Séparation de sources [GR13], débruitage d'images [LCHR19], reconnaissance de visages [WYG + 09], super-résolution d'images [JWHY08], extraction de caractéristiques locales d'images [MBP + 08], segmentation de mouvement [RTVY08]

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Notre contribution [Thèse de Huanzhang Fu] :

→ **Application de ces principes à la classification d'images en proposant une représentation parcimonieuse reconstructive et discriminative des images**

- Terme discriminant dans la fonction objectif de la représentation parcimonieuse
- Apprentissage d'un dictionnaire reconstructif et discriminatif

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Modèle de 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 :

$$y = Dx$$

Choix de la solution correspondant au minimum de la norme l^0

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

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Reconstructive and Discriminative Sparse Representation for Visual Object Categorization (RDSR_VOC) :

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 $1 \leq i \leq M$.

Fonction objectif de la représentations parcimonieuse :

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

Terme discriminant

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Codage parcimonieux : Sequential Forward Sparse Coding (SFSC)

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 $X = D_{\Gamma^t}^+ Y$.

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Algorithme complet

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 $\omega_k : \{i | 1 \le i \le N, x_T^k(i) \neq 0\}$ 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.

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Expérimentations

Jeu de données Simplicity (1000 images, 10 catégories)

Validation croisée à 4 échantillons,

2446 descripteurs d'images, 60 atomes

Incrustation vidéo

Descripteurs basés sur une représentation parcimonieuse des images

Résultats (taux de classification)

	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%

Publié dans : ICIG 2009, BMVC 2011

Incrustation vidéo

Fusion multimodale

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 $E_k = \{C_k^{j^1}, C_k^{j^2}, ..., C_k^{j^M}\}$

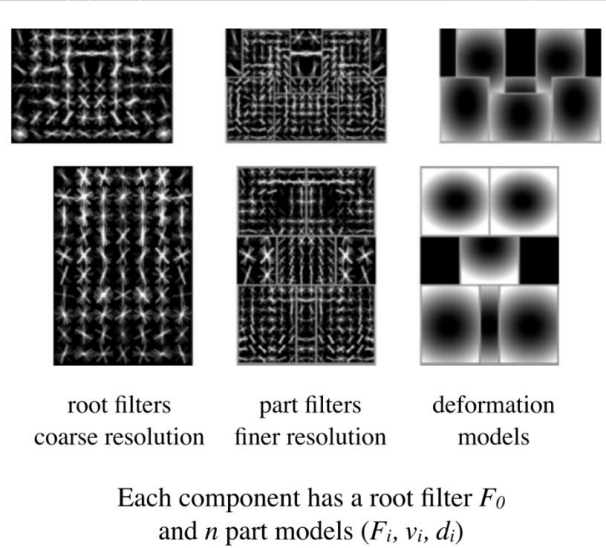
- For the number of experts $n = 2$ to M
  - For each concept $k = 1$ to K
    1. Select the first n experts from E_k : $E_k^n = \{C_k^1, C_k^2, ..., C_k^n\}$
    2. Select the first n weights from W_k : $W_k^n = \{\omega_k^1, \omega_k^2, ..., \omega_k^n\}$
    3. For $j = 1$ to n : Normalise $\omega_k^{j'} = \omega_k^j / \sum_{i=1}^n \omega_k^i$
    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

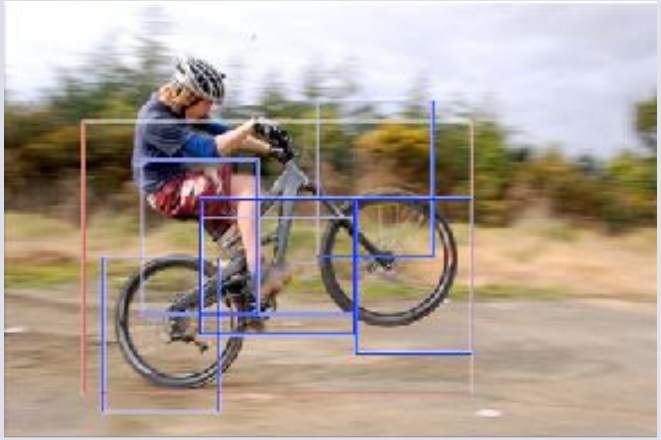
Incrustation vidéo

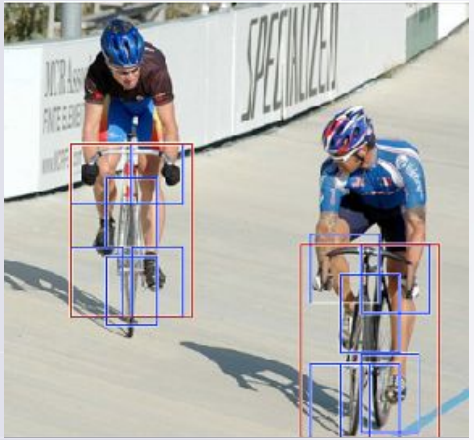
Apprentissage faiblement supervisé de DPM

Modèles à parties déformables (DPM) [FGMR10] :

➔ Ex : Modèle de vélo à 2 composants







Incrustation vidéo

Apprentissage faiblement supervisé de DPM

DPM complètement supervisé [FGMR10] :

Modèle : 1 racine + plusieurs parties

- **Racine :** couvre l'objet annoté
- **Parties :** peuvent se déplacer dans l'objet (déformation)
  - Position de la racine : position donnée par l'annotation (boîte englobante)
  - Position des parties : inconnues (variables latentes)

Descripteurs : pyramide HOG

Classifieur : latent SVM (LSVM)

Incrustation vidéo

Apprentissage faiblement supervisé de DPM

DPM faiblement supervisé [PL11] :

!! Pas d'annotation au niveau de boîtes englobantes

- **Position de la racine :** inconnue

  ➜ initialisée aléatoirement (recouvrement > 40% de l'image)

  ➜ rapport de forme : moyenne de ceux dans l'ensemble d'apprentissage

- **Position des parties :** inconnues (variables latentes)

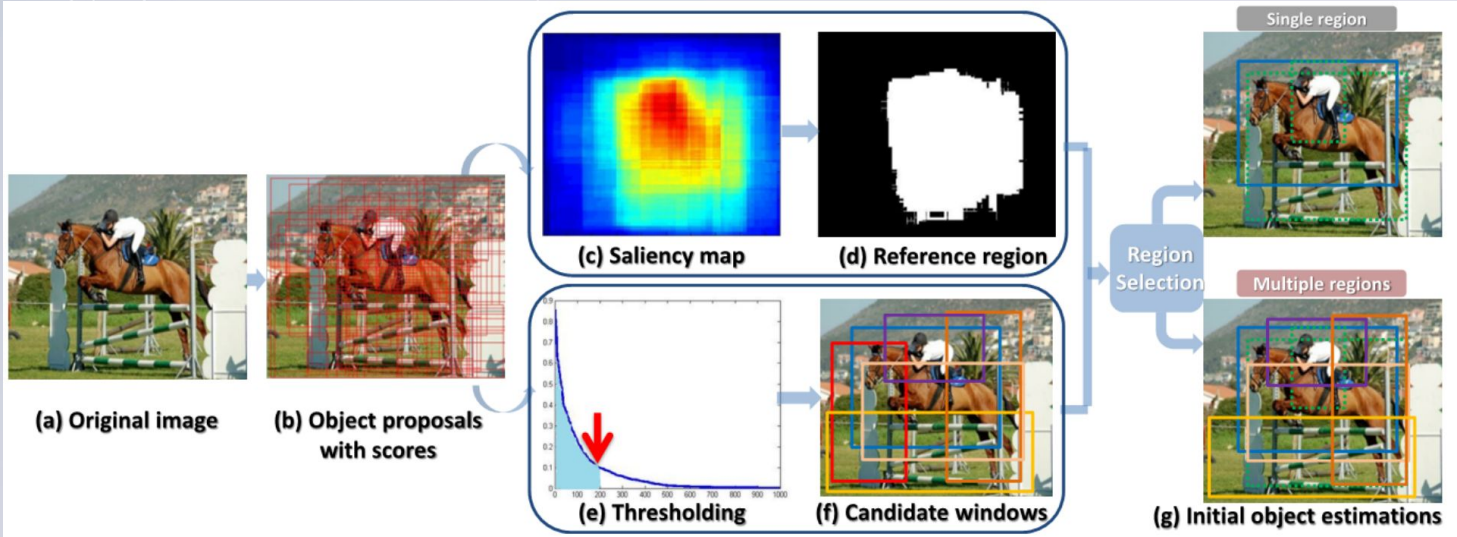
Descripteurs : pyramide HOG

Classifieur : latent SVM (LSVM)

Incrustation vidéo

Apprentissage faiblement supervisé de DPM

Notre proposition [Thèse de Yuxing Tang] : estimation des positions initiales

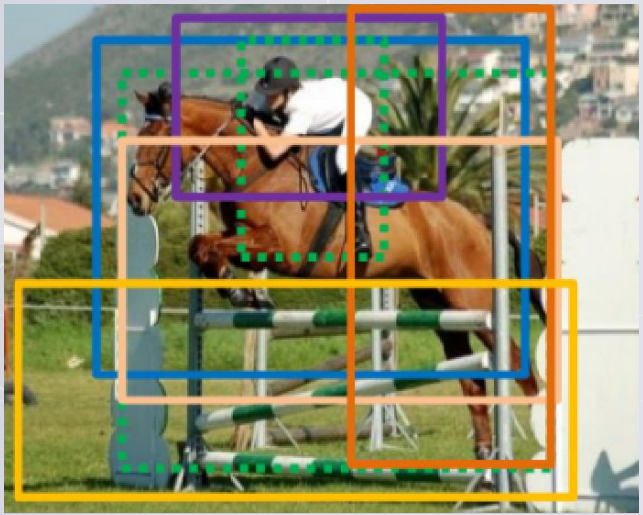


Incrustation vidéo

Apprentissage faiblement supervisé de DPM

Notre proposition : apprentissage des classes latentes par classification de régions

Labels uniquement au niveau de l'image globale : cheval / personne



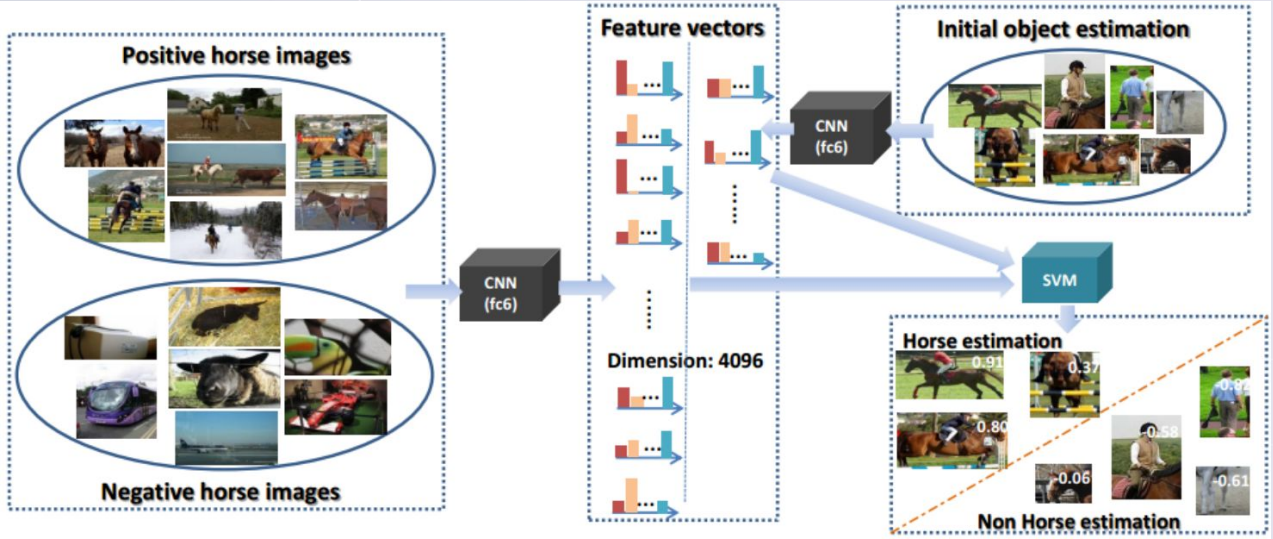
Quelle fenêtre candidate contient un cheval ?

Quelle fenêtre candidate contient une personne ?

Incrustation vidéo

Apprentissage faiblement supervisé de DPM

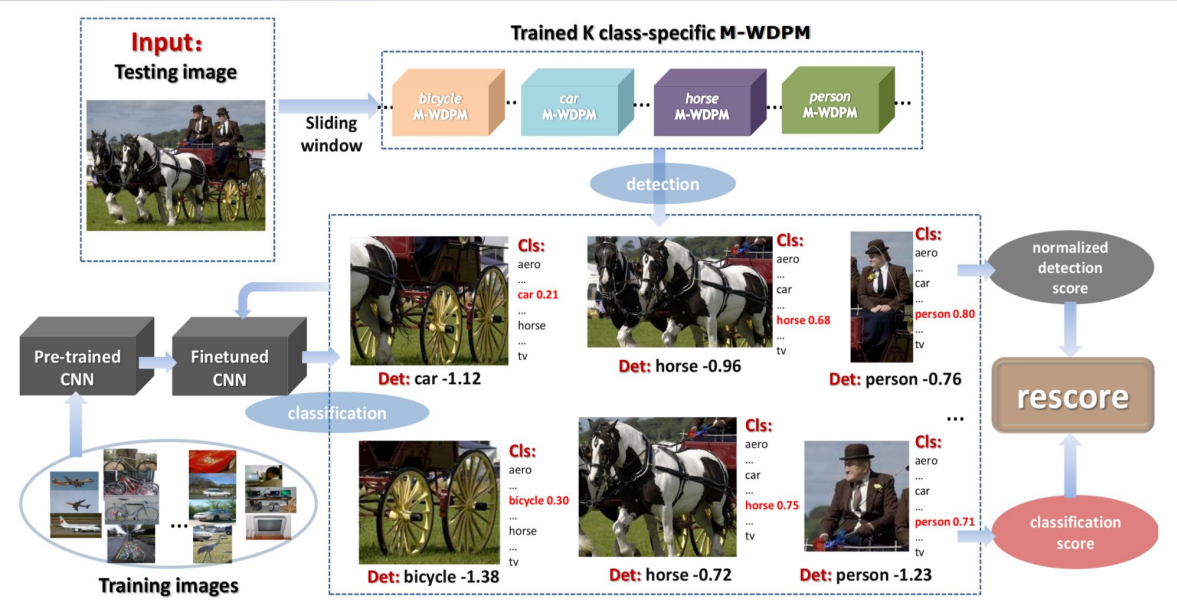
➜ Utilisation de classifieurs binaires au niveau de l'image globale pour attribuer un score aux régions candidates



Incrustation vidéo

Apprentissage faiblement supervisé de DPM

Utilisation du contexte : combinaison de classifieurs et de détecteurs



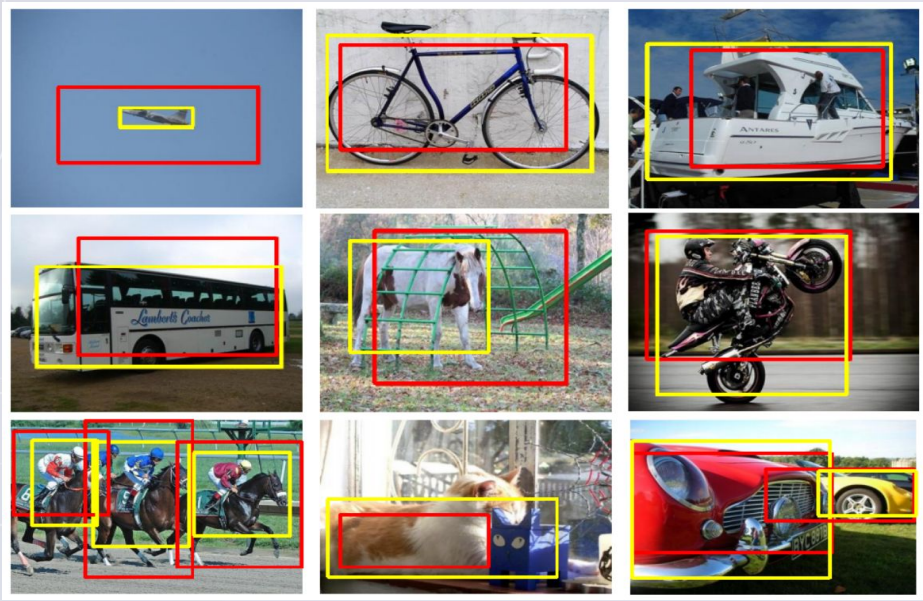
Incrustation vidéo

Apprentissage faiblement supervisé de DPM

Ajustement des boîtes englobantes : élargissement-contraction basé sur l'énergie des contours

Avant post-traitement

Après post-traitement



Incrustation vidéo

Apprentissage faiblement supervisé de DPM

Résultats: jeu de données PASCAL VOC 2007, 9963 images, 20 catégories d'objets. Précision moyenne (en %) sur l'ensemble de test

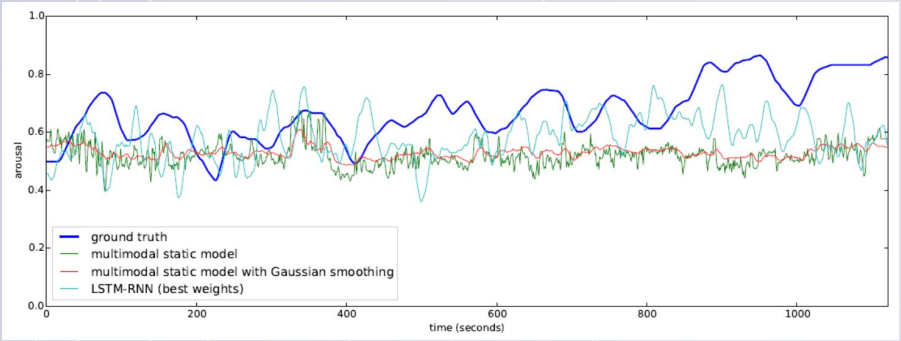
method / class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
our M-WDPM-HOG	34.1	41.5	15.2	10.0	8.8	36.5	40.8	31.5	4.6	23.1	9.4	24.2	29.8	42.5	9.1	14.5	18.3	11.2	32.1	14.3	22.6
our M-WDPM-deep	38.2	38.4	17.5	15.8	9.5	38.1	39.4	32.0	3.5	26.4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

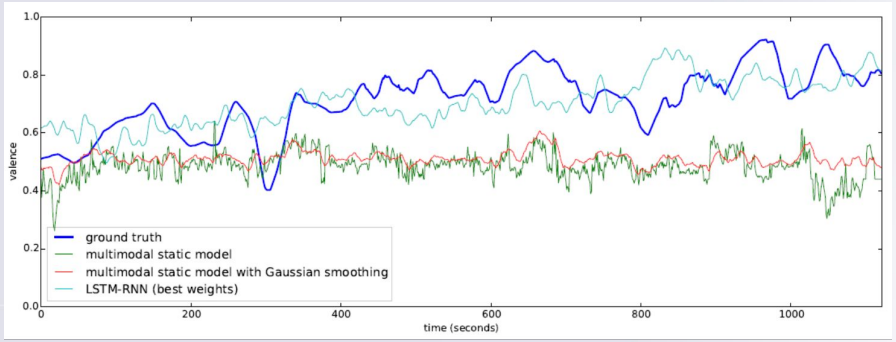
Publié dans : ICIP 2014, IEEE Transactions on Multimedia (2017)

Incrustation vidéo

Modèle de prédiction spatio-temporel

Expérimentations :





Incrustation vidéo