Stage de fin d'année

Reconnaissance et détection des interactions humaines

Présenté par :
OUALI Yassine

Composition du Jury :
M. Alain Mérigot
Mme. Christine Parey
M. Abdelhafid Elouardi
M. Omar Hammami

Encadrement :
M. Bertrand Luvison
Mme. Astrid Orcesi

M2 SETI - Année universitaire 2017-2018

PLAN

1. Introduction
2. Motivation
3. Création du dataset
4. Détection des interactions
5. Conclusion & perspectives

À propos du stage :

CEA TECH

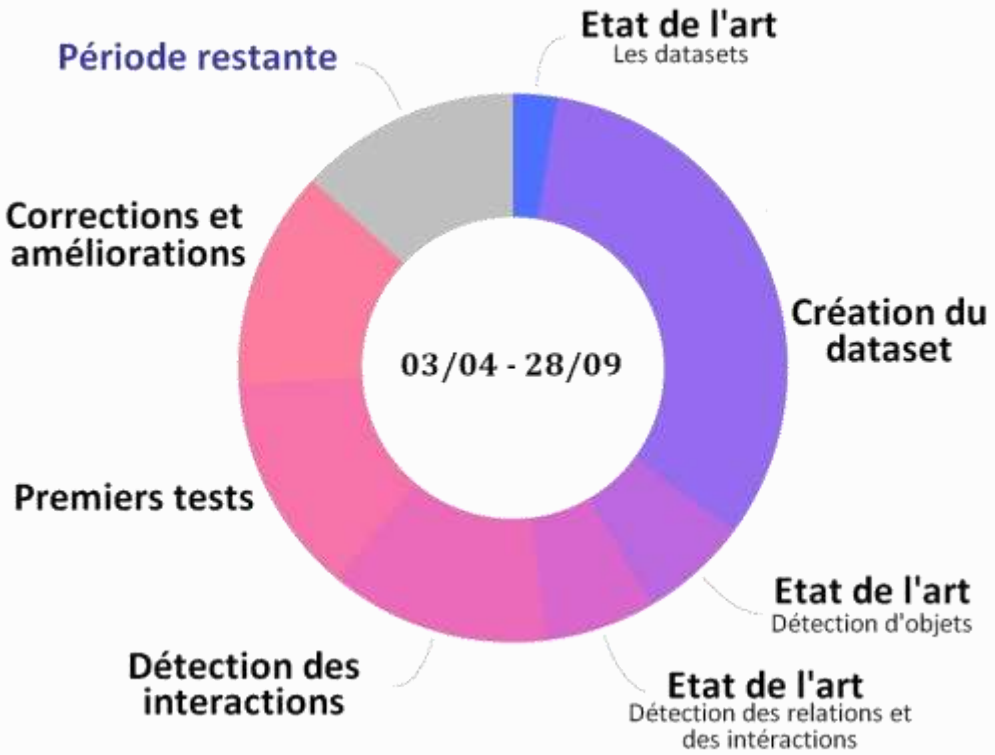
DRT
Direction de la Recherche Technologique

LIST
Département Ingénierie Logiciels et Systèmes

DIASI
Département Intelligence Ambiante et Systèmes Interactifs

LVIC
Equipe : Analyse de scène

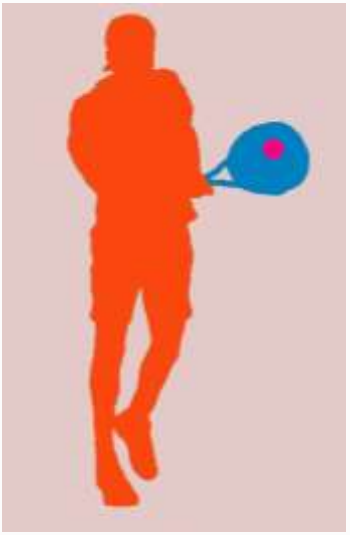
Déroulement du stage :



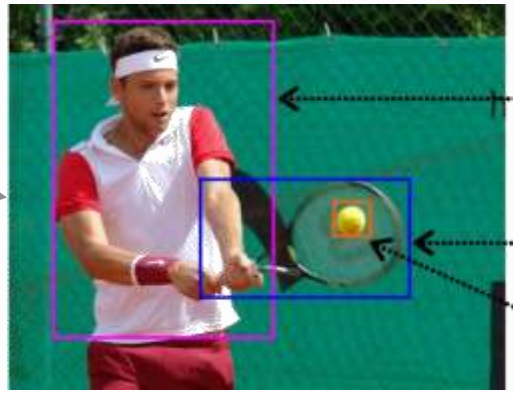
Aller au-delà d'une detection simple

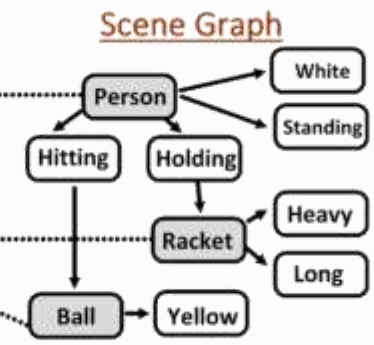
Estimation de pose

Détection d'objets



Segmentation





Triplets

Ball yellow
Racket heavy
Racket long
Person white
Person standing
Person holding racket
Person hitting ball
Person holding racket

Interactions

Relations visuelles

Verbes - Comparaison - Propositions - Action - Spatiale


person wear shirt


elephant taller that person


motorcycle with wheel


person kick ball


person on top of road

Interactions

Verbes - Actions

Intéractions Homme - Homme

Intéractions Homme - Objet

Deux types d'interactions

Interactions Homme-Objet

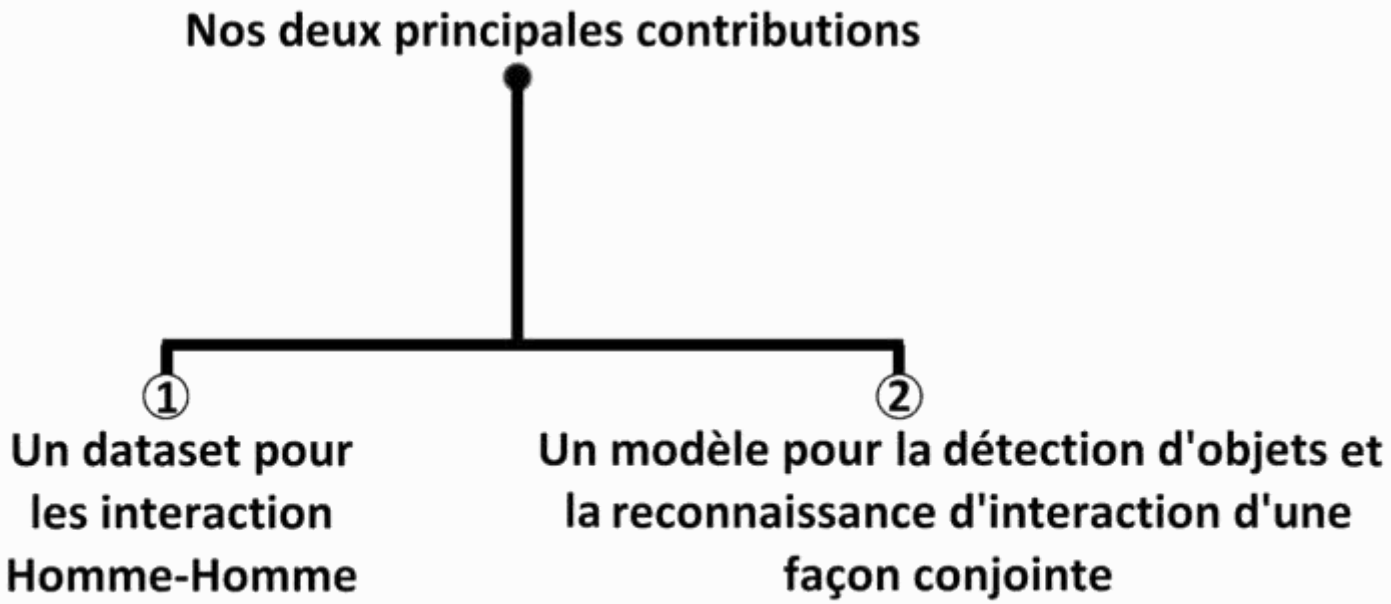

Eating an apple


Cutting an apple

Interactions Homme-Homme


Feeding

Highfive


Nos deux principales contributions
1
Un dataset pour les interaction Homme-Homme
2
Un modèle pour la détection d'objets et la reconnaissance d'interaction d'une façon conjointe

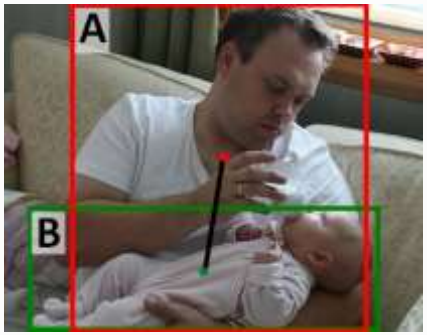
Représentation des interactions

Homme-Homme

Triplets : <*sujet*, verb, *objet*>

Action commutative

< *A*, highfive, *B* >
< *B*, highfive, *A* >



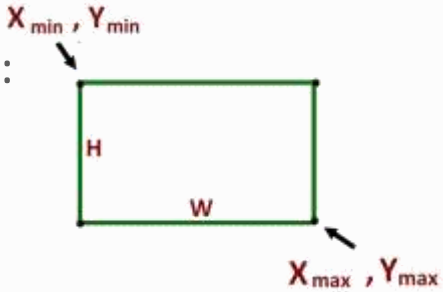
Action non commutative

< *A*, feeding, *B* >
< *A*, holding, *B* >

Action individuelle

<*A*, saluting, ~ >
<*B*, saluting, ~ >

A et B sont des objets annotés :
[Coordonnées, ID, catégorie]



Comparaison des datasets d'interactions

	Datasets	Taille	Verbes	Triplets	Annotation des objets
Homme-Homme	ShakeFive2	153 vidéos	8	✗	✗
	Human Interaction Images	1972 images	9	✗	✗
	K3HI	312 poses	8	✗	✗
	SBU Kinect Interaction	171 vidéos	8	✗	✗
	BIT-Interaction	400 vidéos	7	✗	✗
	UT-Interaction	60 vidéos	6	✗	✗
	TV Human Interaction	350 vidéos	5	✗	✗
Homme Objet	HICO	47K images	117	✗	✗
	HICO-DET	47K images	117	✓	✓
	VCOCO	10K images	27	✓	✓

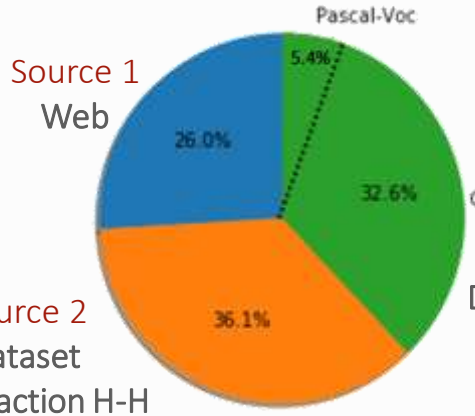
→ Création d'un dataset des interactions H-H

- ~ 5000 Images
- 25 Verbes
- Annotations des objets et des interactions

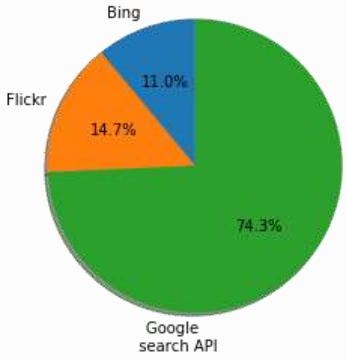
arguing	bending	choking	fistbump	holding
handshake	helping	highfive	interviewing	kissing
nursing	patting	playing	haircut	pulling_pushing
giving	saluting_waving	feeding		
kicking	punching_boxing	hugging		
talking	throwing	thumbsup	dancing	

Création du dataset

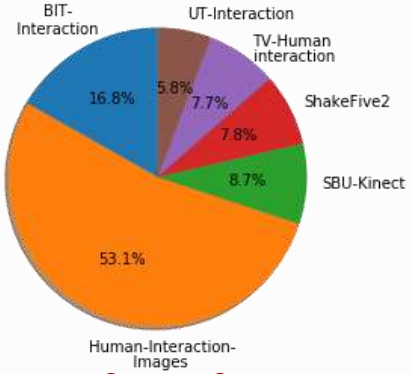
Collection des images



3 Sources



Source 1



Source 2

1

Annotation des objets

2

Annotation de 80 catégories d'objets

- Pré-annotation avec le *Detectron*
- Correction manuelle

Création du dataset

Outil d'annotation

Design d'outil d'annotation pour ajouter les triplets.

Vidéo de démonstration

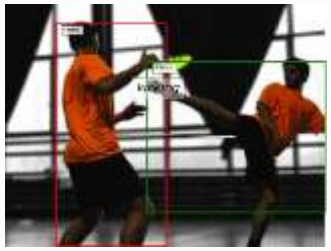
3

Annotation des interactions

Ajout des interactions en forme de triplets

4

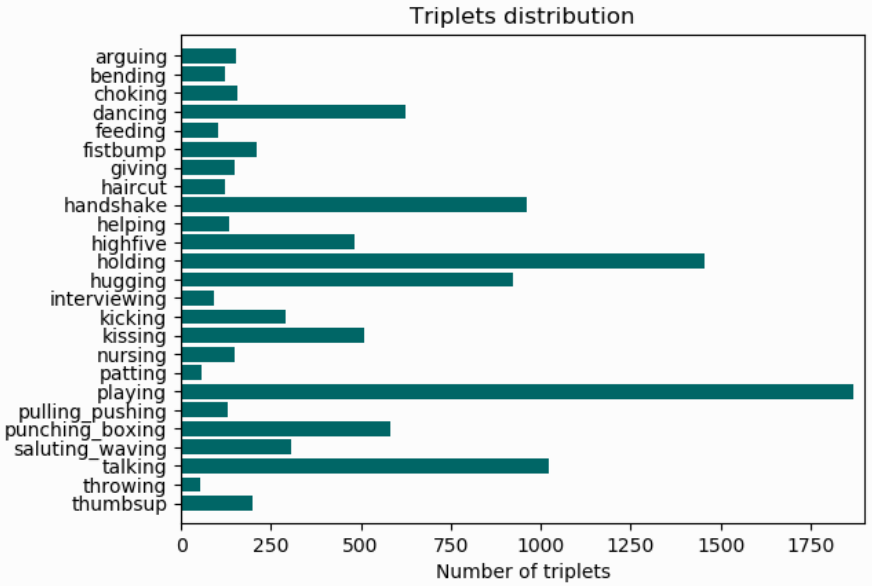





Quelques statistiques

	Datasets	Type	Taille	Verbes	# Objets	Triplets	Ann. O & S	Ann. Objets
Homme-Homme	ShakeFive2	Vidéos	153	8	-	✗	✗	✗
	Human Interaction Images	Images	1972	9	-	✗	✗	✗
	SBU Kinect Interaction	Vidéos	171	8	-	✗	✗	✗
	BIT-Interaction	Vidéos	400	7	-	✗	✗	✗
	UT-Interaction	Vidéos	60	6	-	✗	✗	✗
	TV Human Interaction	Vidéos	350	5	-	✗	✗	✗
	Notre dataset	**Images**	**5.5K**	**25**	**78**	✓	✓	✓
Homme Objet	HICO-DET	Images	47K	117	80	✓	✓	✗
	VCOCO	Images	10K	27	80	✓	✓	✓

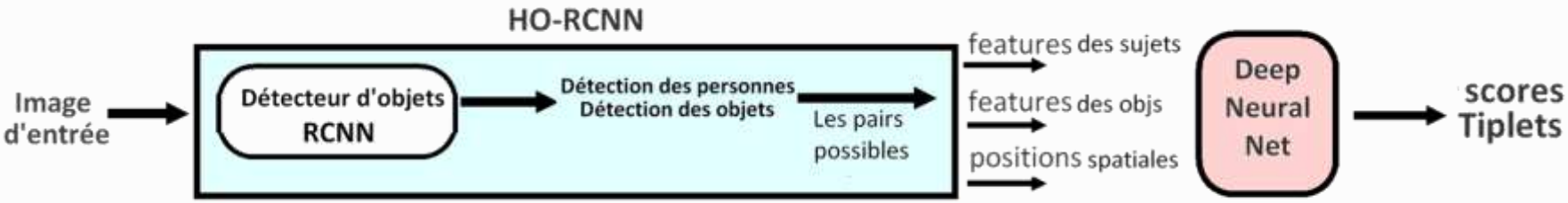
- Taille totale: **862 MB**,
- Nombre d'images: **5417,**
- Nombre de verbes: **25,**
- Nombre de triplets: **10854**
- Nombre de segmentations: **17639,**
- Nombre de boites: **37713,**
- Nombre d'instances de personnes: **21721.**



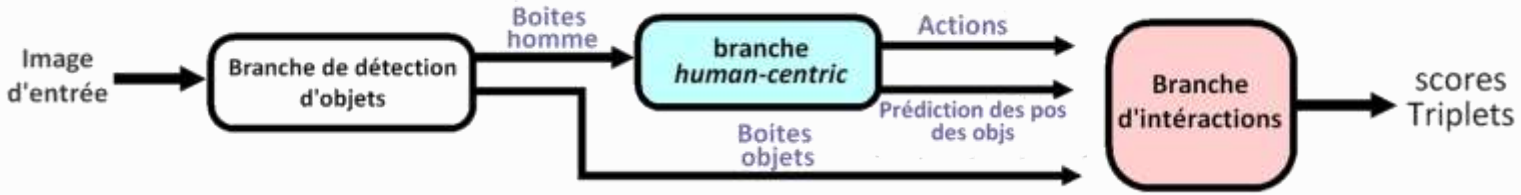
Détection des interactions

État de l'art

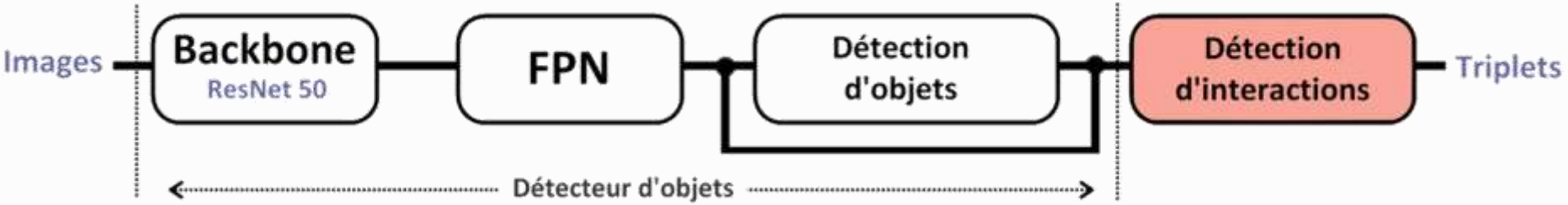
Méthode 1 : **Learning to Detect H-O Interactions**, Chao et al, [2018].



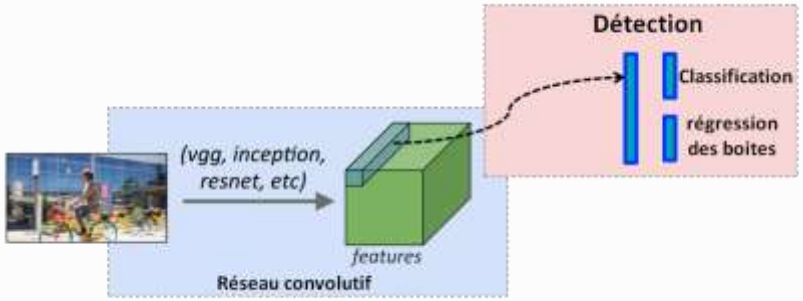
Méthode 2 : **InteractNet**, Gkioxari et al [2018].



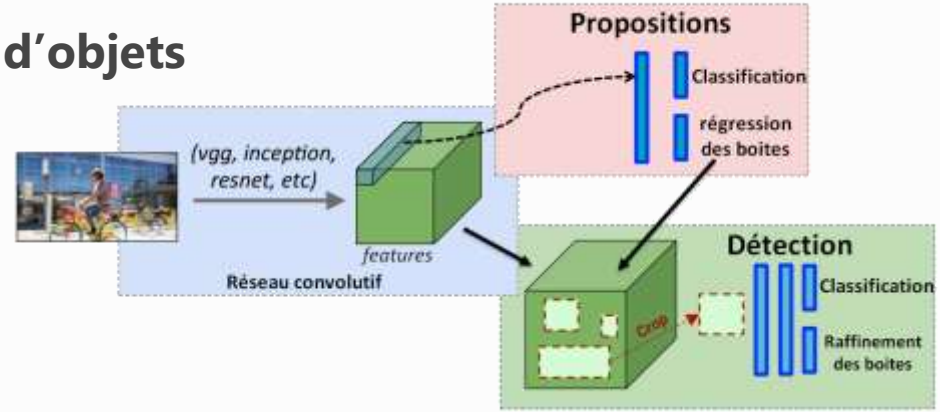
Notre méthode



Détecteurs d'objets

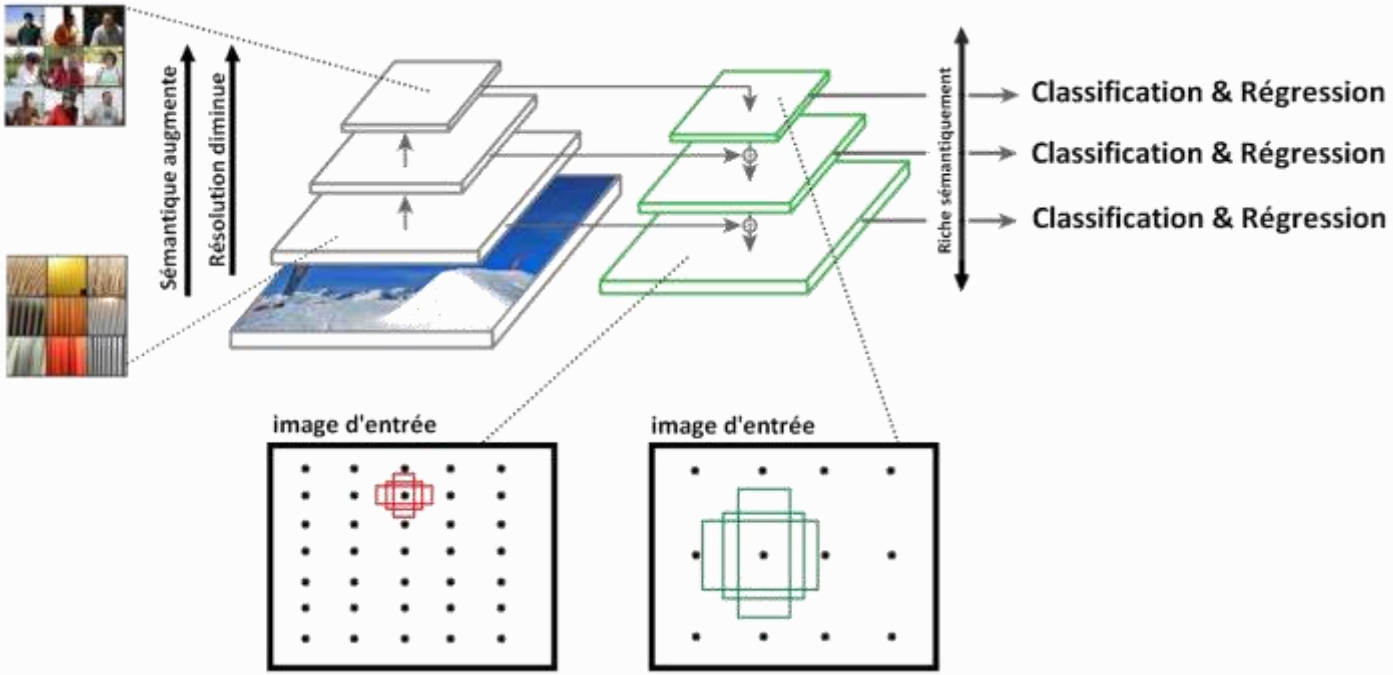


Détection directe

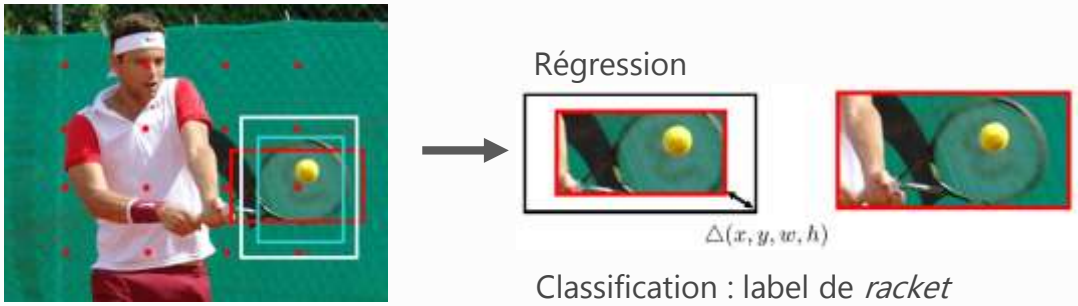


Détection raffinée

Détecteur de choix : **RetinaNet** Lin et al (2017).

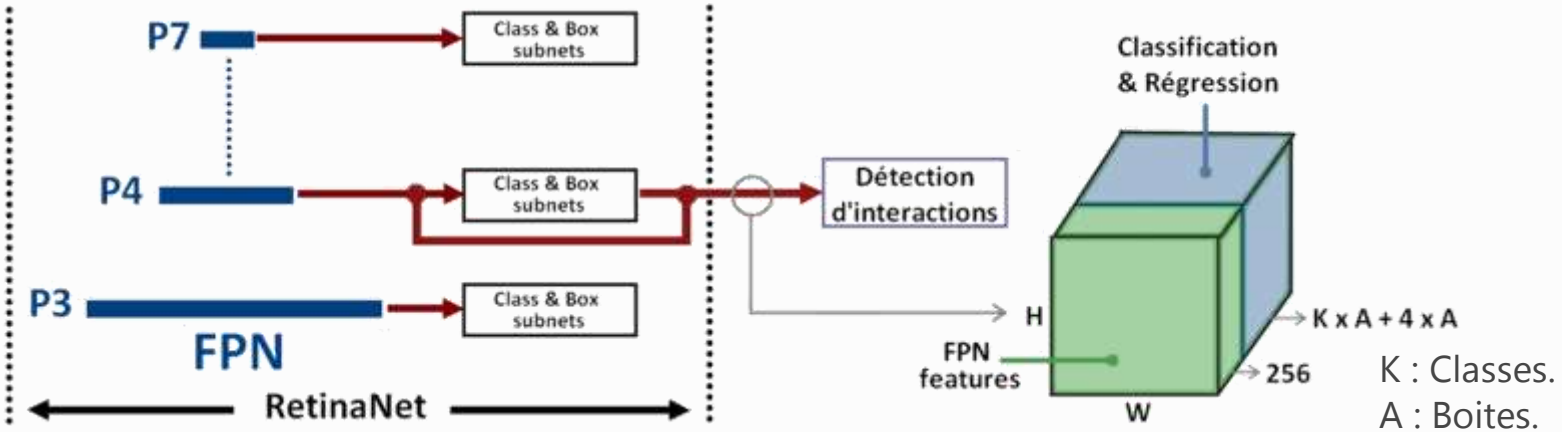


Exemple de détection



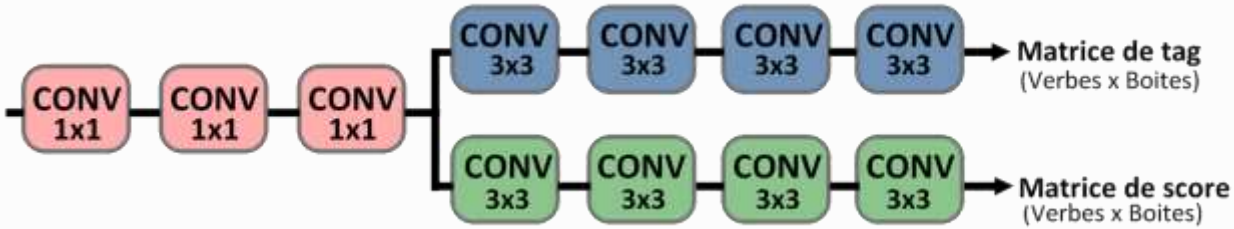
Classification : label de *racket*

Modèle proposé



K : Classes.
A : Boites.

Module de detection d'interactions



Les sorties : Tags & Scores

Les triplets:

<B1, V2, B2>
<B3, V2, B2>
<B3, V1, B1>
<B5, V4, B4>

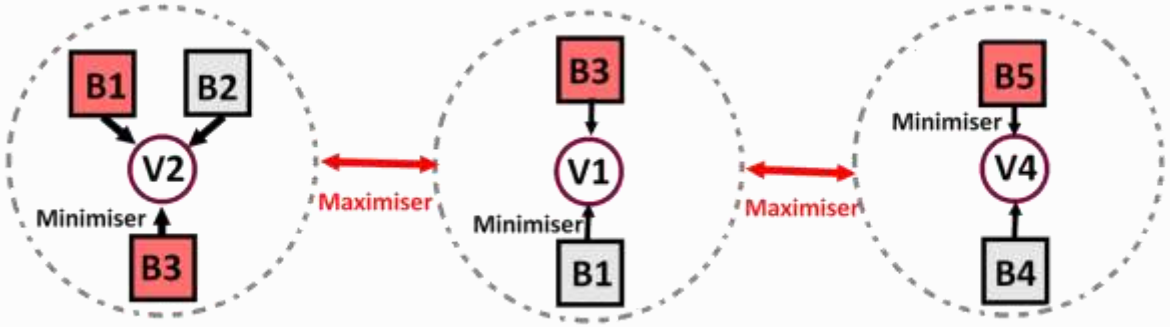
Matrice de Score:

	V1	V2	V3	V4
B1	0	1	0	0
B2	0	0	0	0
B3	1	1	0	0
B4	0	0	0	0
B5	0	0	0	1

Matrice de Tag:

	V1	V2	V3	V4
B1	γ	α	0	0
B2	0	α	0	0
B3	γ	α	0	0
B4	0	0	0	β
B5	0	0	0	β

- Sujets



Objectives

Tags

- Maximiser les distances entre les boites qui sont pas en interactions,
- Minimiser les distances entre les boites en interactions.

Scores

- Classifier les actions faites par les sujets.

Inférence :

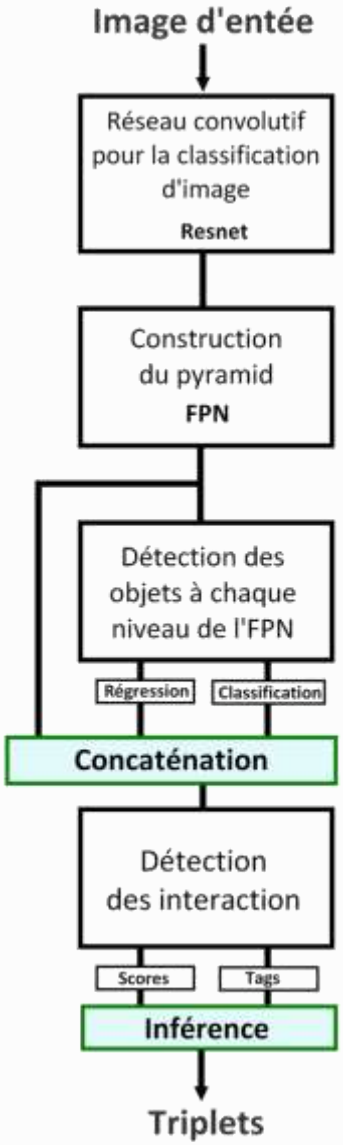


Image d'entée
Réseau convolutif pour la classification d'image
Resnet
Construction du pyramid
FPN
Détection des objets à chaque niveau de l'FPN
Régression
Classification
Concaténation
Détection des interaction
Scores
Tags
Inférence
Triplets

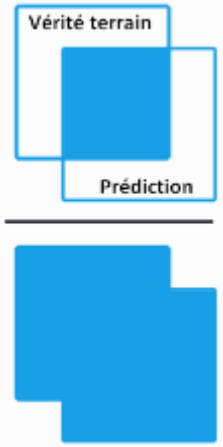
Métriques d'évaluation

<*sujet*, verbe, *objet*>

AP_{agent} : Le couple sujet verbe est correct.

AP_{role} : Le triplet sujet, verbe, objet est correct.

Pour le sujet / objet $IoU = \frac{\text{Vérité terrain} \cap \text{Prédiction}}{\text{Vérité terrain} \cup \text{Prédiction}} > 0{,}5$

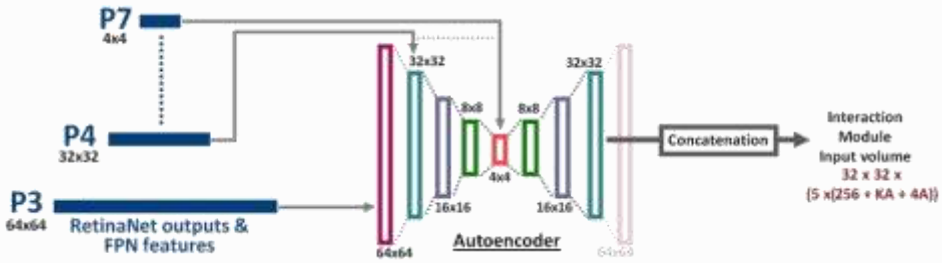


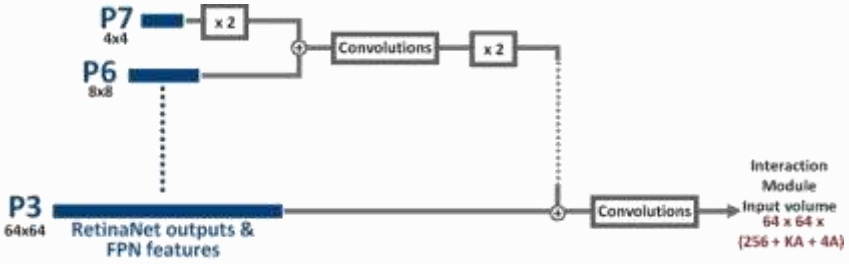
Expériences

Modèle					Dataset	Interactions	AP agent	AP rôle
	FPN	Module d'inter.	L_{tag}	L_{score}				
InteractNet					VCOCO	H-O	**69.2**	**40.0**
Notre modèle	1 niveau	1×1 & 3×3	Marge	CS	VCOCO	H-O	21.64	-
Notre modèle	1 niveau	1×1 & 3×3	Marge	CS	VCOCO	H-O	21.07	-
Notre modèle	1 niveau	1×1 & 3×3	Exp	CS	VCOCO	H-O	22.45	-
Notre modèle	1 niveau	1×1 & 3×3	Exp	CS	VCOCO	H-O	29.92	-
Notre modèle	1 niveau	1×1 & 3×3	Marge	L2	Notre dataset	H-H	8.5	-
Notre modèle	1 niveau	1×1 & 3×3	Marge	CS	Notre dataset	H-H	25.51	-

Futures expériences

- Utilisation de plusieurs niveaux pour la détection des interactions





- Tester des différents hyper-paramètres et optimiseurs.
- Modifier les fonctions de pertes pour mieux satisfaire les objectives d'apprentissage

Perspectives

- Avoir de meilleurs résultats sur la détection des interactions : H-H et H-O
- Publication du dataset avec les premiers benchmarks

A court terme : Détection de l'ensemble des interactions

Datasets	Taille	Verbes	# Objets	Triplets	Anno. S&O	Anno. Objets
Notre dataset	5.5K	25	78	✓	✓	✓
VCOCO	10K	27	80	✓	✓	✓
All interactions	**15.5K**	**52**	**80**	✓	✓	✓

A long terme : Détection de l'ensemble des relations dans les scènes visuelles avec des modèles plus fiables et moins complexe.

Conclusion

De nombreux défis restent à relever avant que nous puissions réaliser la vision des machines capables de comprendre le monde visuel et d'interagir avec nous à travers le langage naturel.

Références

- Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. ECCV, 2014.
- Ross Girshick Kaiming He Piotr Dollár Tsung-Yi Lin,Priya Goyal. Focal loss for dense object detection. CVPR, 2017.
- Piotr Dollàr Georgia Gkioxari, Ross Girshick and Kaiming He. Detecting and recognizing human-object interactions. CVPR, 2018.
- Xieyang Liu Huayi Zeng Yu-Wei Chao, Yunfan Liu and Jia Deng. Learning to detect human-object interactions, WACV, 2018.
- Chen Sun Menglong Zhu Anoop Korattikara-Alireza Fathi Ian Fischer Zbigniew Wojna Yang Song Sergio Guadarrama Kevin Murphy Jonathan Huang, Vivek Rathod. Speed/accuracy trade-offs for modern convolutional object detectors. CVPR, 2017.
- Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. arXiv, 2015.

Merci de votre attention

Annexe

Fonctions de pertes

Fonction de perte :

$$L_{tag} = L_{pull} + L_{push}$$

Moyenne des tags :

$$\bar{h}_n = \frac{1}{K} \sum_k h_k(x_{nk})$$

Tag pull:

$$L_{pull} = \frac{1}{N} \sum_n \sum_k (\bar{h}_n - h_k(x_{nk}))^2$$

Tag push:

$$L_{push\ exp} = \frac{1}{N^2} \sum_n \sum_{n'} \exp\left\{ \frac{1}{2\sigma^2} (\bar{h}_n - \bar{h}'_n) \right\}^2$$

$$L_{push\ margin} = \sum_n \sum_{n'} \max\left(0, \|\bar{h}_n - \bar{h}'_n\|\right)$$

Fonction pour les scores:

$$L_{score\ L2} = \sum_k (S_k - \hat{S}_k)^2$$

$$L_{score\ CS} = \sum_k -(1 - S_k) \log\left(1 - \hat{S}_k\right) - S_k \log\left(\hat{S}_k\right)$$

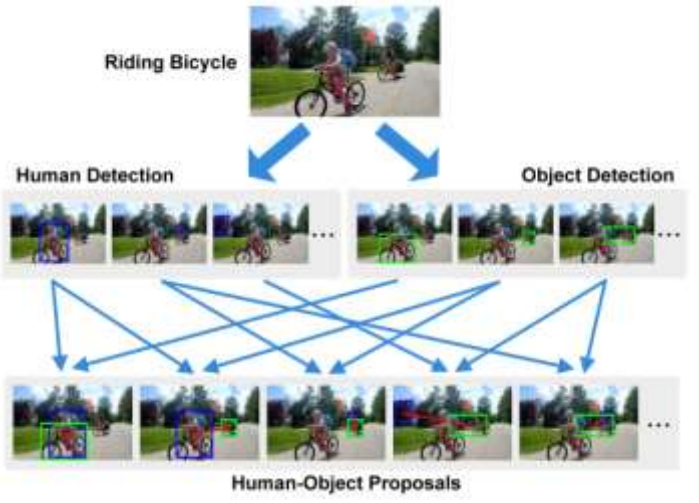
Annexe
Etat de l'art

Learning to Detect Human-Object Interactions – [M1]

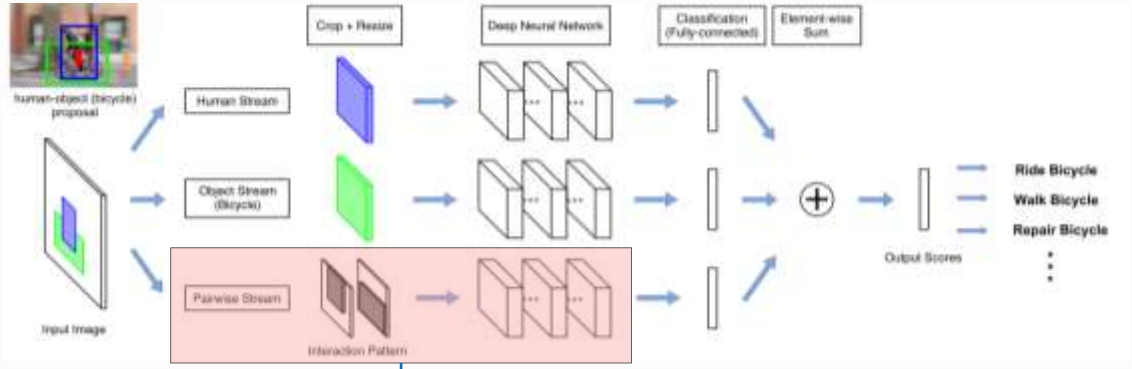
Générer des propositions de pairs H-O

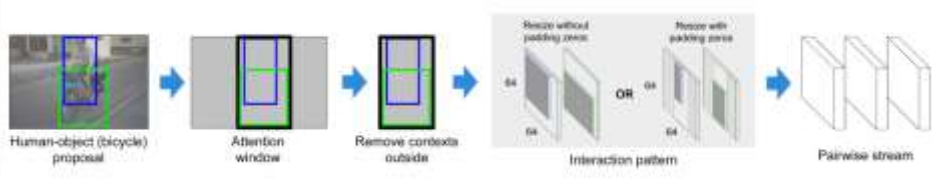
↓

Classification des pairs H-O



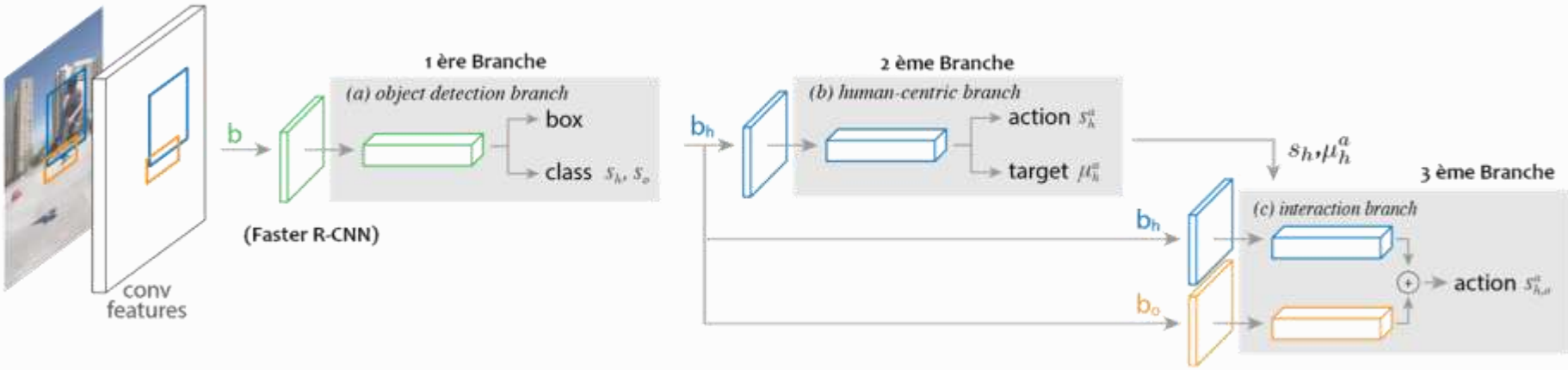
Modèle :





Annexe
Etat de l'art

Detecting and Recognizing Human-Object Interactions (*InteractNet*)



1ère Branche : les scores s_h et s_o de classification

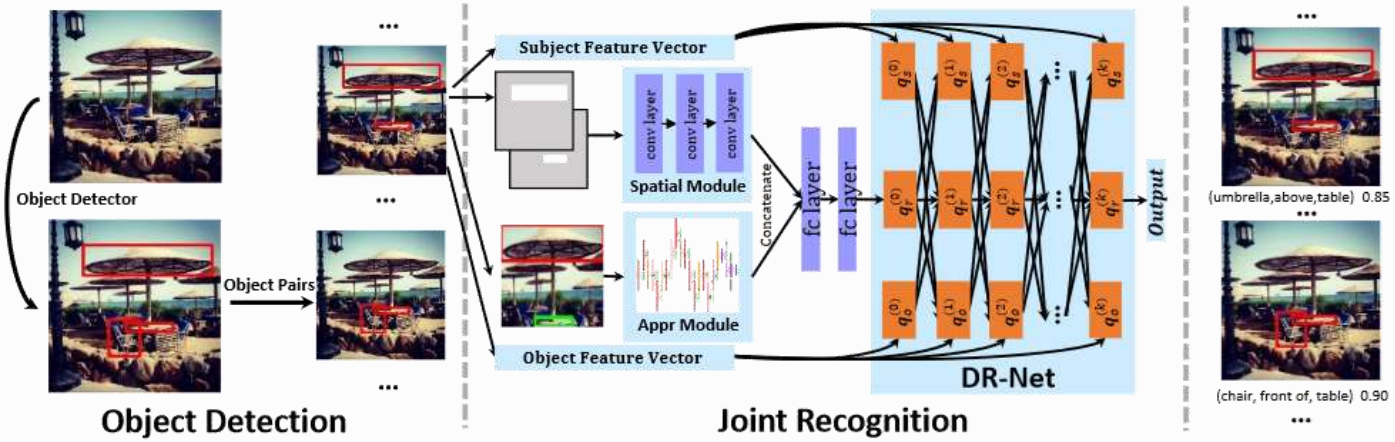
2ème Branche :

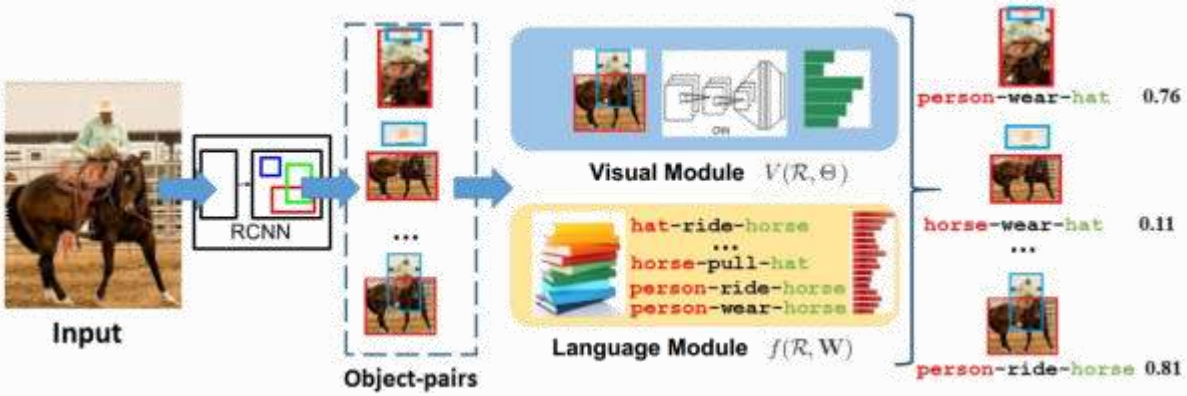
- *Reconnaissance d'actions*: scores s_h^a, action $a \ -> \ b_h$.
- *Prédiction de positionnement de l'objet en question*: prédire une densité sur l'emplacement possible de l'objet.

3ème Branche : Détection des triplets <action a pour les bbox o et h>

Annexe
Etat de l'art

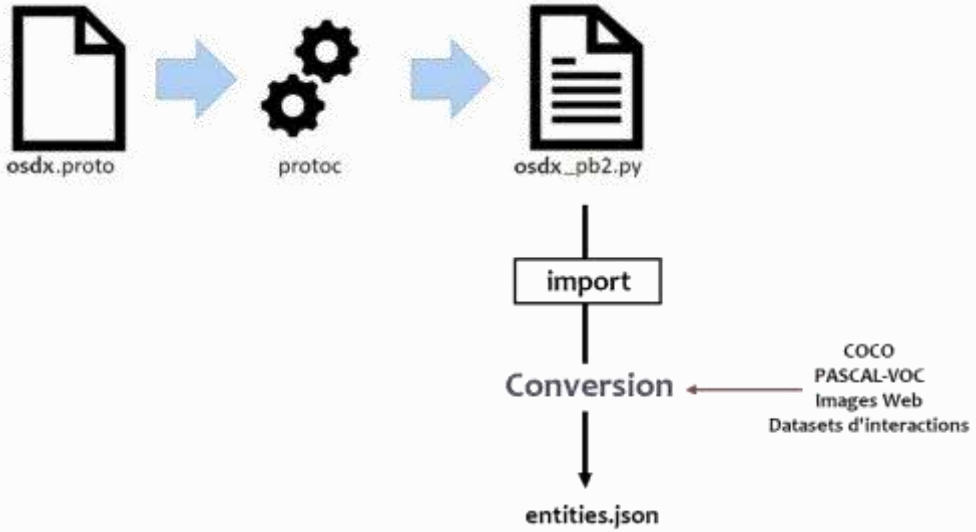
Detection des relations





Annexe
Dataset

4 – Conversion des annotation en format **osdx**.

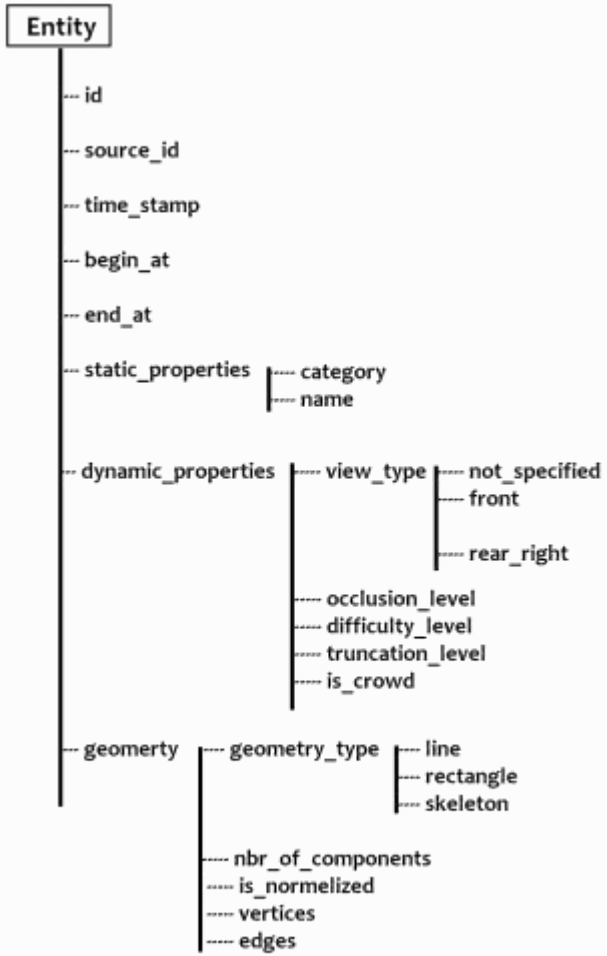


5 – Annotation des bbox avec **ATOMiC**:

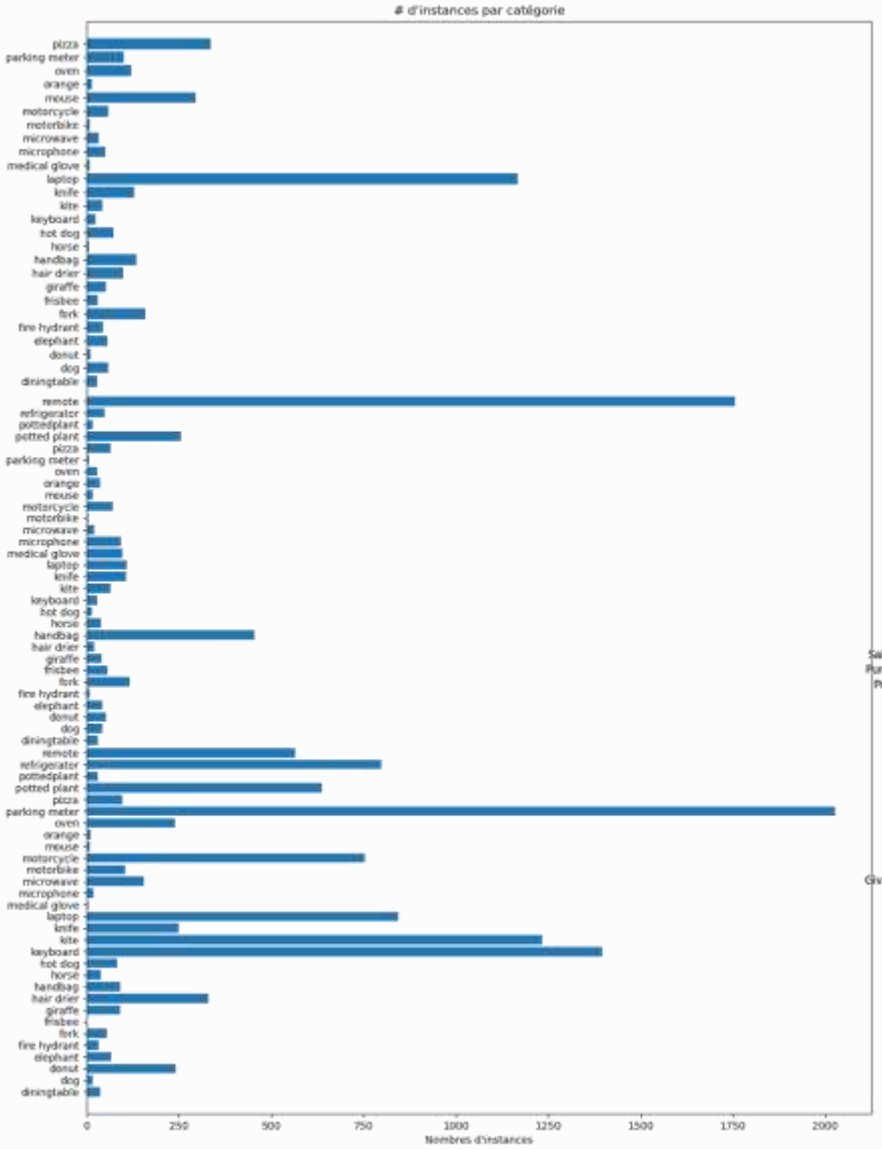
Les images de web : 1539,
Les images ∈ datasets des interactions : 2184,
Les 80 classes de COCO,

-> Pré-annotation avec le Detectron
-> Correction manuel des annotations (~2 min / img)

osdx.proto

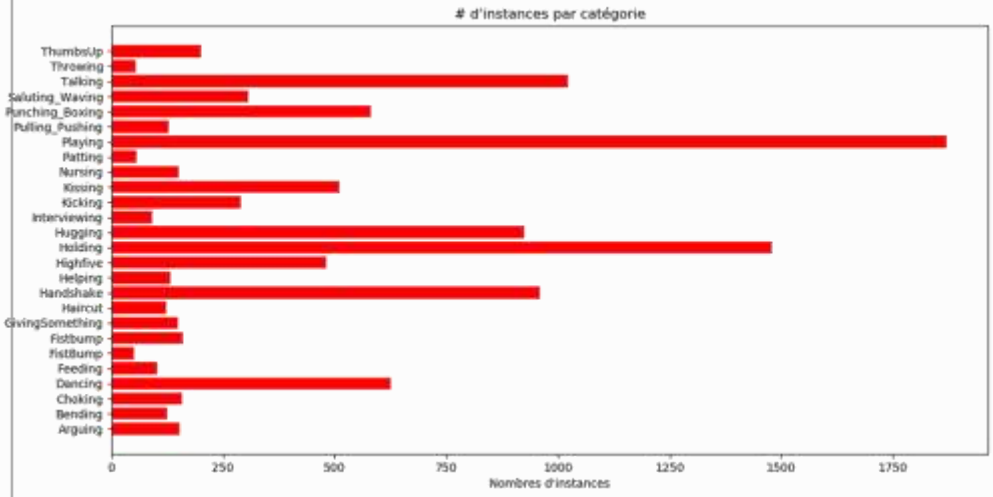


Annexe
Dataset



- Total size : 862 Mo
- Total images : 5417
- Average image size (px) : 908 x 688
- Minimum image size (px) : 213 x 142
- Maximum image size (px) : 7500 x 5000

- Total number of interactions : 10878
- Total number of segmentations : 21449
- Total number of bbox : 39393
- Total number of person instances : 22264



Annexe

Exemples d'images

Choking

Patting

Boxing

Highfive

Holding

Saluting

Playing

Feeding

Giving

Talking

Hugging

Dancing

Haircut

Kicking

.....