Classification instantanée de mouvements de foules dans des vidéos

Antoine BASSET*, Patrick BOUTHEMY, Charles KERVRANN

Inria, Centre Rennes – Bretagne Atlantique
Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
Antoine.Basset@inria.fr, Patrick.Bouthemy@inria.fr, Charles.Kervrann@inria.fr

Résumé – Nous nous intéressons à la classification de mouvements de foules denses dans des vidéos. Contrairement à la plupart des méthodes d'analyse de foule, la méthode proposée ne nécessite pas d'intégration temporelle ou d'apprentissage ; elle produit des cartes denses de classification instantanée, *i.e.* calculées à partir de deux images successives et évaluées en chaque point mobile de l'image.

Abstract – This paper deals with the problem of recognizing motion classes in densily crowded scenes. In contrast to most crowd analysis methods, the proposed algorithm does not require any time integration nor learning phase, and delivers a frame-based pixel-wise crowd motion classification.

1 Introduction

La plupart des méthodes d'analyse de foules dans des vidéos (*e.g.* [15]) s'intéressent au suivi de piétons [7, 10], à la détection de comportements anormaux [5, 8, 12] et à la classification de trajectoires [14, 17]. Dans [16] et [2] les auteurs se concentrent sur les mouvements cohérents et dominants. Wang *et al.* [14] et Zhou *et al.* [17] cherchent à extraire les chemins les plus empruntés. À notre connaissance, seuls Hu *et al.* [7] et Solmaz *et al.* [13] ont étudié la classification des mouvements de groupes. Les premiers déterminent des types de mouvement en partitionnant des champs de vecteurs 4D (positions et vitesses dans l'image) à l'aide de critères de proximité et de similarité [7]. Les seconds évaluent par advection les trajectoires d'un ensemble de particules et étudient, aux lieux d'accumulation, les configurations dynamiques associées [13].

Pour catégoriser les mouvements de foules dans des vidéos, nous proposons une nouvelle approche utilisant seulement une paire d'images, alors que les méthodes usuelles d'analyse de foules reposent sur des intervalles de temps plus longs (au moins une dizaine d'images). Il s'agit alors d'analyses de cuboïdes spatiotemporels [5, 8, 11], de tracklets [17] et plus généralement de trajectoires [2, 10, 13, 14, 16]. Contrairement à ces approches, notre méthode repose sur des modèles affines de mouvement 2D, estimés sur une paire d'images. Il n'y a ni intégration temporelle, ni calcul de trajectoire, ni phase d'apprentissage, ni réglage fin d'un paramètre. Nous catégorisons les mouvements apparents de la foule, c'est-à-dire dans le plan de l'image, en fonction d'un cadrage, supposé fixe.

Nous supposons que le mouvement apparent d'un groupe de piétons peut localement être représenté par un mouvement de translation, d'homothétie ou de rotation, et nous déduisons huit classes. Les mouvements « homothétiques » correspondent au rapprochement des piétons (classe *convergence*) ou à leur éloignement (*divergence*). On distingue les mouvements de rotation *directs* et *indirects*, et quatre directions principales de translation pertinentes dans l'image : *nord*, *ouest*, *sud* et *est*. Nous avons évalué notre méthode sur des séquences synthétiques et des images réelles. Elle peut s'étendre à l'analyse de groupes d'entités en mouvement (véhicules, animaux...).

L'article est organisé comme suit. La section 2 introduit les modèles de mouvement que nous estimons à partir d'une collection de fenêtres. Nous décrivons ensuite la méthode de sélection du meilleur modèle en chaque point et la méthode de classification des mouvements de foule observés. Dans la section 3, nous présentons et commentons nos résultats expérimentaux. Enfin, nous concluons dans la section 4.

2 Analyse des mouvements de foule

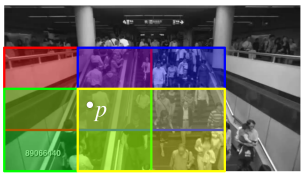
Notre méthode se décompose en quatre étapes : (i) détection des points mobiles, (ii) estimation des modèles de mouvement candidats dans une collection de fenêtres, (iii) sélection du modèle optimal en chaque point selon un critère de maximum de vraisemblance (MV), (iv) classification par un arbre de décision utilisant des votes majoritaires. Pour la première étape, nous avons recours à l'algorithme de détection de mouvement [3], basé sur des champs de Markov à états mixtes [3]. On note Ω le domaine de l'image et $\mathcal{S} \subset \Omega$ l'ensemble des points mobiles détectés.

2.1 Estimation des modèles de mouvement

Nous ne considérons que des modèles paramétriques de mouvement 2D. En chaque point $p = (x, y) \in \mathcal{S}$, le flot optique $w(p)$

*La thèse d'A. Basset bénéficie du soutien de la Région Bretagne.

(a) Grande échelle (100%)



(b) Échelle moyenne (50%)

(c) Petite échelle (25%)

FIG. 1 : Ensemble des fenêtres contenant un point p (séquence *Escaliers*). En (a) les fenêtres jaune et rouge sont tronquées pour s'inscrire dans l'image.

est approché par le vecteur de vitesse $w_\theta(p)$ défini par :

$$w_\theta(p) = \underbrace{\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}}_{A} \begin{pmatrix} x \\ y \end{pmatrix} + \underbrace{\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}}_{B}. \quad (1)$$

Nous notons $\theta = (a_1, a_2, a_3, a_4, b_1, b_2)^T$ le vecteur de paramètres du modèle de mouvement. Pour caractériser les huit classes de mouvement de foule, seuls trois modèles « sous-affines » de mouvement sont nécessaires : translation, homothétie et rotation. Ils correspondent respectivement aux matrices A suivantes, telles que décrites dans [6] :

$$A_T = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \; A_H = \begin{pmatrix} a_1 & 0 \\ 0 & a_1 \end{pmatrix}, \; A_R \begin{pmatrix} 0 & a_2 \\ -a_2 & 0 \end{pmatrix}. \quad (2)$$

Le vecteur B correspond à la vitesse de l'origine du repère. Selon le modèle, nous n'estimons que deux (translation) ou trois (homothétie et rotation) coefficients :

$$\theta_T = (b_1, b_2)^T, \; \theta_H = (b_1, b_2, a_1)^T, \; \theta_R = (b_1, b_2, a_2)^T. \quad (3)$$

Comme nous ne connaissons pas à l'avance le support adéquat pour estimer les modèles de mouvement, nous considérons une collection $\mathcal{F}$ de fenêtres de tailles variées (25%, 50% et 100% des dimensions de l'image). Pour chaque taille, le taux de recouvrement est de 50%, de sorte qu'un point p appartient à trois ou quatre fenêtres de cette taille (voir la figure 1).

Nous estimons les trois modèles de mouvement (3) dans chaque fenêtre par la méthode robuste multirésolution [9]. La procédure IRLS (Iteratively Reweighted Least Squares) utilisée associe à tout point p un poids évaluant son influence dans l'estimation. Un point de poids proche de 1 contribue de manière significative à l'estimation des paramètres. On note $\theta_{k,i}$ le vecteur de paramètres du modèle $k \in \{T, H, R\}$ estimé dans $F_i \in \mathcal{F}$, et $\mathcal{X}_{k,i}$ l'ensemble de ses points significatifs obtenu par seuillage sur les poids.

La conformité d'un point p à un modèle de paramètres $\theta_{k,i}$ est donnée par la différence d'images déplacées :

$$\varepsilon(p, \theta_{k,i}) = I_{t+1}(p + w_{\theta_{k,i}}(p)) - I_t(p), \quad (4)$$

où I_t désigne les intensités de la t-ième image, et $w_{\theta_{k,i}}(p)$ la vitesse de p déduite de (1) pour $\theta_{k,i}$. La conformité correspond à une valeur de ε proche de 0. Pour chaque modèle de mouvement et dans chaque fenêtre F_i, nous estimons le vecteur de paramètres $\theta_{k,i}$ [9] et la variance empirique $\sigma^2_{k,i}$ définie par :

$$\sigma^2_{k,i} = \frac{1}{|\mathcal{X}_{k,i}|} \sum_{p \in \mathcal{X}_{k,i}} \varepsilon^2(p, \theta_{k,i}), \quad (5)$$

où $|\mathcal{X}_{k,i}|$ est le cardinal de $\mathcal{X}_{k,i}$.

Soit $\mathcal{F}(p) \subset \mathcal{F}$ l'ensemble des fenêtres contenant le point p, et $\mathcal{M}(p)$ l'ensemble associé des modèles de mouvement candidats en p. Avec la distribution de fenêtres décrite précédemment, nous estimons 33 modèles de mouvement candidats par point (30 aux bords de l'image).

2.2 Sélection des modèles de mouvement

Parmi tous les modèles proposés en un point p, nous choisissons le plus pertinent suivant un critère MV. En supposant que les résidus $\varepsilon(p, \theta_{k,i})$ sont indépendants et distribués selon une loi normale centrée, la vraisemblance $L(p, \theta_{k,i})$ du modèle de paramètres $\theta_{k,i}$ évaluée dans un voisinage petit $\mathcal{V}_p$ centré en p, s'exprime de la façon suivante :

$$L(p, \theta_{k,i}) = \prod_{q \in \mathcal{V}_p} \frac{1}{\sqrt{2\pi\sigma^2_{k,i}}} \exp - \frac{\varepsilon^2(q, \theta_{k,i})}{2\sigma^2_{k,i}} \quad (6)$$

où $\sigma^2_{k,i}$ est la variance définie en (5).

Le modèle de mouvement optimal s'obtient ainsi :

$$\hat{\theta}(p) = \underset{\theta_{k,i} \in \Theta(p)}{\arg\max} L(p, \theta_{k,i}), \quad (7)$$

avec $\Theta(p)$ l'ensemble des paramètres associé à $\mathcal{M}(p)$. Si ce modèle de mouvement optimal est une homothétie (resp. rotation), nous le retenons si et seulement si $|a_1|$ (resp. $|a_2|$) est plus grand qu'un certain seuil τ (fixé à 10^{-3}). Si ce seuil n'est pas atteint, nous attribuons à p un mouvement de translation, ce qui revient à poser $a_1 = 0$ (resp. $a_2 = 0$). Ceci permet de limiter l'apparition d'homothéties ou de rotations de faible amplitude. On peut également considérer ce seuillage comme une pénalisation des modèles d'homothétie et de rotation, plus complexes que le modèle de translation.

2.3 Classification

À partir des trois modèles de mouvement considérés, et en fonction du signe des coefficients a_1, a_2, b_1, b_2, ou de certaines combinaisons, nous pouvons construire huit classes de mouvement de foule. Par commodité, nous les représentons par des couleurs, comme indiqué dans le tableau 1 : $\mathcal{C}_T = \{$●, ●, ●, ●$\}$, $\mathcal{C}_H = \{$●, ○$\}$, $\mathcal{C}_R = \{$●, ●$\}$ et $\mathcal{C} = \mathcal{C}_T \cup \mathcal{C}_H \cup \mathcal{C}_R$.

Nous commençons par calculer une carte de classes préliminaire $\mathcal{C}_{pr} = \{c_{pr}(p) \in \mathcal{C} \mid p \in \mathcal{S}\}$, que l'on régularisera ensuite. Elle est obtenue à partir des paramètres estimés du modèle de mouvement sélectionné en (7). Par exemple, considérons le point p tracé sur la figure 1. Il appartient à 11 fenêtres, donc 33 modèles de mouvement donnés par (3) et leurs vraisemblances (6) sont évalués. La vraisemblance la plus élevée est obtenue avec un mouvement d'homothétie, donc $c_{pr}(p) \in \mathcal{C}_H$. De plus, $a_1(p) = -0.0044 < -10^{-3}$, d'où $c_{pr}(p) =$ ●.

TAB. 1 : Critères définissant chaque classe

Modèles de mouvement		Classes de mouvement de foule	
		Directions	Critères
Translation	●	Nord	$b_1+b_2>0, b_1-b_2<0$
	●	Ouest	$b_1+b_2<0, b_1-b_2<0$
	●	Sud	$b_1+b_2<0, b_1-b_2>0$
	●	Est	$b_1+b_2>0, b_1-b_2>0$
Homothétie	●	Convergence	$a_1<0$
	○	Divergence	$a_1>0$
Rotation	●	Indirecte	$a_2<0$
	●	Directe	$a_2>0$

Pour régulariser $\mathcal{C}_{pr}$, nous utilisons un arbre de décision à deux niveaux. Pour chaque point p, les décisions sont prises par vote majoritaire dans une fenêtre carrée centrée, notée $\mathcal{P}_p$; pour retrouver les mouvements de groupes étendus, on prend pour côté de $\mathcal{P}_p$ 25% de la largeur de l'image. La première décision consiste à choisir en chaque point s'il est en mouvement de translation, convergence, divergence, rotation directe ou rotation indirecte. Si la translation est retenue, un second vote majoritaire dans $\mathcal{P}_p$ permet de choisir sa direction parmi les quatre possibles.

3 Résultats expérimentaux

L'algorithme proposé est massivement parallèle. De plus, les calculs de vraisemblance et les votes sont optimisés par l'utilisation d'images intégrales introduites dans [4]. Avec un ordinateur à 4 cœurs de fréquence 2.3 GHz et 8 Go de mémoire vive cadencée à 1.6 GHz, une paire d'images est traitée en 1 à 9 secondes en fonction du nombre de points mobiles.

Nous avons évalué notre méthode avec trois ensembles de vidéos : un premier de synthèse [1], et deux autres réels collectés dans [11, 13]. Tous les exemples présentés ont été traités

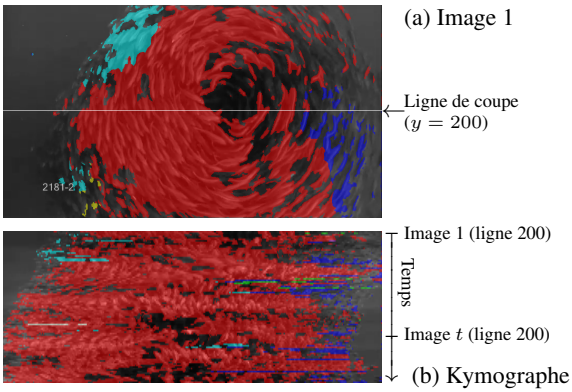
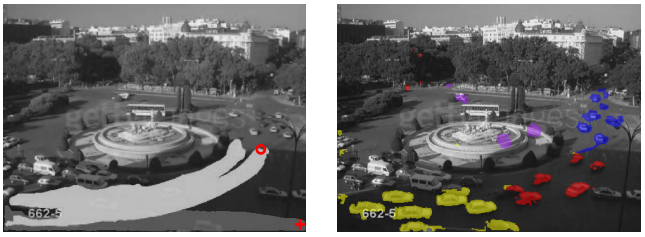


FIG. 2 : Stabilité temporelle (sur 100 images) illustrée sur *Banc de poissons*. (b) représente l'évolution temporelle de la classification de la ligne indiquée en (a). Les zones sombres du kymographe correspondent à des mouvements non détectés. Parmi les points détectés en mouvement, 86.9% sont bien classés (●). La plupart des erreurs (●, ●, ●, ●) sont dues à la faible courbure de la rotation.



(a) Résultat tiré de [13] (b) Notre classification

FIG. 3 : Comparaison entre (a) la méthode de Solmaz *et al.* et (b) la nôtre. Dans *Rond-point*, des voitures circulent du coin inférieur gauche vers le coin supérieur droit. Dans (a), resp. (b), une file (+), resp. un mouvement de translation vers l'est (●), est détectée tout en bas de l'image, et un virage (○), resp. un mouvement de rotation directe (●) en haut. Notre méthode détecte une autre translation vers le nord (●), et une voiture roulant vers l'ouest (●). Nous observons quelques détections parasites sur la fontaine et les arbres.

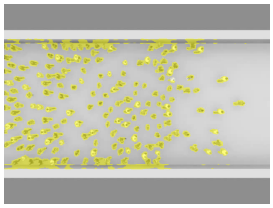
avec les mêmes paramètres : on utilise 3 tailles de fenêtres d'estimation (25%, 50% et 100% des dimensions de l'image), les côtés de $\mathcal{V}_p$ et $\mathcal{P}_p$ mesurent respectivement 3 pixels et 25% de la largeur de l'image, et τ vaut 10^{-3}.

Les figures 2 à 5 présentent des situations variées en termes d'angle de vue, de vitesses, de densité de piétons, et de classes de mouvement. Les résultats respectifs sont commentés dans les légendes de ces figures. Les performances sont souvent très bonnes (figure 4), aux imperfections de segmentation près (figure 2). Malgré l'utilisation de seulement deux images consécutives, la classification est assez stable temporellement, comme le montre le kymographe de la figure 2, qui présente l'évolution de la classification sur 100 images.

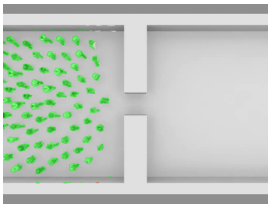
Enfin, pour comparer notre méthode avec celle de Solmaz *et al.*, nous avons analysé la vidéo *Rond-point*, étudiée dans [13]. La figure 3 met en évidence les différences : la méthode [13] classe les trajectoires évaluées sur un long intervalle de temps autour de points stationnaires, alors que notre algorithme produit une *classification instantanée à l'échelle du pixel*. Ainsi, même si elle peut être sujette à quelques instabilités temporelles, notre méthode peut détecter des évènements localisés dans l'image ou se produisant sur un intervalle de temps court.

4 Conclusion

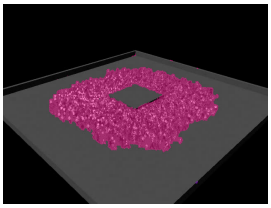
Nous avons proposé une nouvelle méthode de classification de mouvements de foules dans des vidéos, exploitant seulement deux images consécutives. Trois modèles paramétriques de mouvement sont estimés dans une grande collection de fenêtres de différentes tailles. Un critère de maximum de vraisemblance permet de choisir le meilleur modèle en chaque point. La classification finale est obtenue avec un arbre de décision impliquant des votes majoritaires, qui attribue à chaque point mobile l'une des huit classes de mouvement de foule. De plus, même des événements spatialement localisés ou temporellement courts peuvent être retrouvés.



(a) Séquence *Corridor* : marche vers la droite. Résultats pour l'image 1



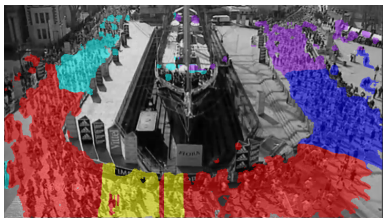
(b) Séquence *Escape* : regroupement vers la sortie. Résultats pour l'image 1



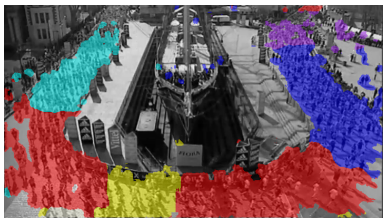
(c) Séquence *Mekkah* : rotation indirecte. Résultats pour l'image 1

	Corridor	*Escape*	*Mekkah*
# ●	0	7	59
# ●	0	0	1601
# ●	0	9	34
# ●	**2392551**	417	78
# ●	15191	**1179719**	3200
# ○	129025	4939	150
# ●	7768	4138	**2447843**
# ●	35782	1885	0
TPR	**92,7%**	**99,0%**	**99,7%**

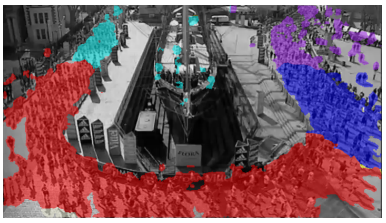
FIG. 4 : Tests sur des séquences de synthèse de 50 images ne contenant chacune qu'une classe de mouvement de foule : (a) *Corridor* correspond à la translation vers l'est (●), (b) *Escape* à la convergence (●) et (c) *Mekkah* à la rotation dans le sens indirect (●). Les cardinaux de chaque classe et les taux de vrais positifs (TPR) sont donnés dans le tableau. Le TPR est calculé comme la proportion de points bien classés parmi ceux détectés en mouvement.



(a) Résultats pour l'image 1



(b) Résultats pour l'image 21



(c) Résultats pour l'image 41

FIG. 5 : La séquence *Virage de marathon*, où les coureurs décrivent un U de la gauche vers la droite. À gauche, ils descendent (●), puis entament un virage à gauche (●). Au milieu du virage, certains points sont étiquetés en translation vers l'est (●) à cause du grand rayon de courbure. Enfin, on retrouve la translation vers le nord (●), mais des points sont mal classés (●) sans doute en raison de quelques individus sur le côté qui marchent vers la gauche.

L'algorithme est rapide et ne requiert ni phase d'apprentissage, ni calcul de trajectoires, ni ajustement fin de paramètres. Les résultats démontrent l'efficacité et la précision de la méthode dans des situations variées.

Nous comptons exploiter cette classification instantanée de mouvements de foules pour élaborer des méthodes de détection d'événements anormaux et de reconnaissance de scénarios dynamiques.

Références

[1] P. ALLAIN, N. COURTY et T. CORPETTI : Agoraset: a dataset for crowd video analysis. *In 1st Int. Work. Pattern Recog. Crowd Anal.*, ICPR'12, Tsukuba, nov. 2012.

[2] A. M. CHERIYADAT et R. J. RADKE : Detecting dominant motions in dense crowds. *J. Selected Topics Sig. Proces.*, 2(4):568–581, août 2008.

[3] T. CRIVELLI, P. BOUTHEMY, B. CERNUSCHI-FRÍAS et J.-F. YAO : Simultaneous motion detection and background reconstruction with a mixed-state conditional Markov random field. *Int. J. Comp. Vis.*, 94 (3):295–316, 2011.

[4] F. C. CROW : Summed-area tables for texture mapping. *ACM SIGGRAPH Comp. Graphics*, 18(3):207–212, juil. 1984.

[5] J. FENG, C. ZHANG et P. HAO : Online learning with self-organizing maps for anomaly detection in crowd scenes. *In 20th Int. Conf. Pattern Recog.*, ICPR'10, Istambul, août 2010.

[6] E. FRANÇOIS et P. BOUTHEMY : Derivation of qualitative information in motion analysis. *Image Vis. Comp.*, 8(4):279–288, nov. 1990.

[7] M. HU, S. ALI et M. SHAH : Learning motion patterns in crowded scenes using motion flow field. *In 19th Int. Conf. Pattern Recog.*, ICPR'08, Tampa, déc. 2008.

[8] L. KRATZ et K. NISHINO : Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. *In IEEE Conf. Comp. Vis. Pattern Recog.*, CVPR'09, Miami Beach, juin 2009.

[9] J.-M. ODOBEZ et P. BOUTHEMY : Robust multiresolution estimation of parametric motion models. *Int. J. Visual Communication Image Representation*, 6(4):348–369, déc. 1995.

[10] M. RODRIGUEZ, S. ALI et T. KANADE : Tracking in unstructured crowded scenes. *In 12th IEEE Int. Conf. Comp. Vis.*, ICCV'09, Kyoto, sep. 2009.

[11] M. RODRIGUEZ, J. SIVIC, I. LAPTEV et J.-Y. AUDIBERT : Data-driven crowd analysis in videos. *In 13th IEEE Int. Conf. Comp. Vis.*, ICCV'11, Barcelone, nov. 2011.

[12] D. RYAN, S. DENMAN, C. FOOKES et S. SRIDHARAN : Textures of optical flow for real-time anomaly detection in crowds. *In 8th IEEE Int. Conf. Adv. Video Sig. Based Surveillance*, AVSS'11, Klagenfurt, août 2011.

[13] B. SOLMAZ, B. E. MOORE et M. SHAH : Identifying behaviors in crowded scenes using stability analysis for dynamical systems. *IEEE Trans. Pattern Anal. Machine Intel.*, 34(10):1–8, 2012.

[14] X. WANG, K. T. MA, G. NG et W. E. L. GRIMSON : Trajectory analysis and semantic region modeling using nonparametric hierarchical Bayesian models. *Int. J. Comp. Vis.*, 95(3):287–312, 2011.

[15] B. ZHAN, D. N. MONEKOSSO, P. REMAGNINO, S. A. VELASTIN et L.-Q. XU : Crowd analysis: a survey. *Machine Vis. Applic.*, 19(5–6):345–357, 2008.

[16] B. ZHOU, X. TANG et X. WANG : Coherent filtering: detecting coherent motions from crowd clutters. *In 12th Eur. Conf. Comp. Vis.*, ECCV'12, Florence, oct. 2012.

[17] B. ZHOU, X. WANG et X. TANG : Random field topic model for semantic region analysis in crowded scenes from tracklets. *In Comp. Vis. Pattern Recog.*, CVPR'11, Colorado Springs, juin 2011.