# Gestion pacifique des foules

**Pierre BERNAS, Guillaume NEE, Philippe DRABCZUK**[1]

[1] **EVITECH – 3 rue Buffon, 91400 ORSAY – +33.820.2008.39**

www.evitech.com

**Résumé** – Nous présentons ici les résultats d'une étude sur la supervision intelligente des foules, dans le cadre de la gestion pacifique des foules. Dans une première partie, nous nous fixons un cadre d'intervention et définissons des objectifs en termes de supervision des foules, puis nous présentons les techniques que nous avons investiguées et utilisées pour réaliser un démonstrateur de vidéo-surveillance intelligente de la foule. Ce document est issu des travaux accomplis dans le cadre du projet RAPID CrowdChecker, soutenu par la DGA en 2010-2012.

## 1. Introduction

Le terme de « Gestion pacifique des foules » se réfère aux techniques de supervision et de maintien de l'Ordre Public en situation de foules, importantes et/ou particulièrement denses, et mobiles, comme par exemple celles que l'on observe dans les grandes infrastructures de transport, dans les grands parcs de loisirs, au sein des grandes villes, ou à l'occasion de grands évènements (les lieux associés seront désignés ici sous le terme de « site »).

Nous proposons de décomposer l'activité de « ***Gestion*** » des foules conduite par les « ***Gestionnaires de foule***[1] » en plusieurs phases : tout d'abord une phase de «*Préparation* », qui consiste en la prise d'actions antérieures à l'arrivée de la foule (cf. méthodes, organisation dans [25]), puis une phase de « *Supervision* », qui consiste à observer et établir un diagnostic de la situation en divers points du site, une fois la foule en place; une phase d'«*Actions* », composée des interactions instantanées avec la foule (via des moyens tel que le son, les éclairages, les panneaux indicateurs, les barrières mobiles, ou l'intervention directe des gestionnaires) ; et finalement une phase de "*Capitalisation*", qui consiste à tirer les leçons des 3 premières phases pour améliorer les processus à l'avenir. Une fois que la foule est partie (ex: dans [25]) on peut analyser les enregistrements (mains courantes, vidéos) afin d'évaluer et d'améliorer les procédés de contrôle des foules.

Les missions essentielles sont de garantir que (1) **chacun arrive sain et sauf à sa destination**, et (2) **en suivant les cheminements autorisés**. Les questions d'ordre individuel (telles que l'identité, le comportement…)[2] ne sont pas pertinentes pour la gestion pacifique des foules.


Cependant, plusieurs études scientifiques techniques ou sociales (ex: [22]) ne considèrent la phase de « *Supervision* » que d'un point de vue purement comportemental, en estimant que l'information sur le comportement individuel observé est un résultat qui peut simplement être obtenu ou déduit de l'observation, et que ce comportement pourra révéler n'importe quelle situation potentiellement dangereuse. Cette approche nous parait trop restrictive.

En effet, nous souhaitons généraliser ici le concept de « *Supervision* », car d'abord (1) nous considérons que le « comportement » relève de la vie privée et non directement du maintien de l'Ordre Public, et (2) :

- **(2.a) Le comportement de la foule peut témoigner d'évènements pas nécessairement**

[1] La terminologie utilisée dans le « Victorian crowd control safety guide » [25] est celle de « Crowd controllers ». En France, les gestionnaires de foules peuvent être rattachés à la Police, aux Forces armées, aux Pompiers ou à d'autres organismes de sécurité (privée).

[2] Par exemple des questions telles que "Qui était là?", "Qui parle avec qui?" ou "De quelle origine/sexe/âge/couleur de peau était ce groupe?" ou "Quelqu'un a-t-il un comportement étrange?"

**visibles des observateurs de la foule.** Par exemple des personnes s'arrêtent et fixent particulièrement un point ou un lieu ; ou fuient subitement en masse, en courant.

- (2.b) Les accidents les plus graves dans une foule peuvent être la conséquence de situations successives relativement banales, non-accidentelles, comme par exemple une accélération conduisant à une poussée diffuse et à des chutes de personnes d'un quai de gare. Aucun comportement individuel particulier n'est en cause. La prise en compte de ces évènements collectifs peut aider à prévenir les accidents similaires.

La considération « 2.a » consiste à utiliser la foule comme un détecteur qui va révéler des incidents non directement observables aux alentours.

La seconde (2.b) appartient à la catégorie désignée comme un « ***signal faible*** » dans le rapport ESRAB [1] en 2006, c'est à dire une suite d'informations, chacune de faible importance, mais dont l'enchaînement est susceptible de conséquences graves. Par exemple, dans une situation où on observerait de façon régulière que des personnes toussent à proximité d'une petite fuite de gaz, premièrement identifiée comme « une odeur », cette « odeur » peut être le début d'une fuite plus importante, qui deviendra responsable ultérieurement de plusieurs évanouissements et chutes par terre. Dans ce cas, la toux répétitive de personnes de la foule au passage de la fuite pourrait être un signal.

Cependant, même s'il est évident que parfois des « *signaux faibles* » surviennent avant un accident (ex: dans une situation ou les flux de personnes utilisent un escalier pour descendre à un endroit où d'autres flux similaires de personnes ne peuvent sortir ou s'échapper à la même vitesse), il n'est ni évident ni prouvé que ces « *signaux faibles* » apparaissent toujours avant un accident (ex: quelqu'un soudain poussé sur la voie depuis un quai par une personne dont le comportement jusque là était irréprochable).

En conclusion, si nous considérons la variété des accidents qui peuvent se produire, il est clair que certains « signaux faibles » sont parfois visibles et précurseurs d'accidents, d'autres ne sont pas visibles, et enfin certains types d'accidents sont sans précurseur observable dans la foule.

Cet article résulte de travaux réalisés dans le cadre du projet RAPID CrowdChecker, soutenu par la DGA, mené par l'entreprise Evitech et le laboratoire de recherche Willow (ENS/CNRS/INRIA) entre octobre 2010 et août 2012.

Nous présentons un outil pour la **gestion pacifique des foules**. Dans un premier temps nous allons détailler nos objectifs en termes de "*Supervision*", puis nous présenterons notre démarche pour développer un démonstrateur de système d'analyse intelligente d'images de la foule.

# 2. Objectifs de supervision des foules

## 2.1 Criticité des accidents

En examinant de près les appels d'offre relatifs aux systèmes de gestion des foules au niveau mondial, au cours des 5 dernières années (ex: pour les sites de transports, musées, villes, lieux de culte, centres commerciaux, notamment en Europe, Moyen-Orient, Amérique du Sud…), nous nous apercevons que **la demande principale exprimée dans les cahiers des charges est le comptage de personnes**. Même si, les termes de sécurité, de densité et de danger sont aussi mentionnés, la fonction la plus demandée est le comptage de personnes : compter des personnes, détecter si un seuil (nombre de présents) d'alarme est dépassé, mesurer le flux, détecter des regroupements,...

Cependant, on ne peut réduire l'observation de la foule au simple comptage. Compter des personnes est certes utile pour détecter la surpopulation d'un site, détecter un risque d'événement dangereux du fait du nombre des présents, ou pour calculer le montant d'un loyer par rapport à la fréquentation du couloir passant devant un magasin. En résumé, compter nous semble utile mais **ne devrait pas être considéré comme la finalité ultime de la supervision des foules**.

Dans un autre domaine, l'examen de la criticité des équipements d'un aéronef dans l'aviation civile a conduit à la création de documents normatifs (DO178 [2]) qui caractérisent la criticité comme **la perte potentielle qui serait occasionnée par la défaillance de l'équipement correspondant.** Les équipements dont la panne provoquerait la perte de l'avion et la mort de tous les passagers ont reçu le niveau le plus critique (niveau A), alors que ceux nécessitant des procédures accomplies par l'équipage ont été classés à un niveau plus bas (D), et ce qui n'affecte pas le vol se situe au niveau E. Les défaillances d'équipements qui conduiraient à la mort ou à des blessures de quelques personnes sont classées dans les niveaux intermédiaires (B/C). La mention d'une matrice de risques apparait aussi dans [25] comme un outil pour l'évaluation d'incidents dans la foule.

En reprenant ces principes, en supervision des foules, les situations les plus critiques (niveau A), c'est-à-dire celles qui devraient être détectées le plus rapidement possible, sont les accidents de groupes où l'on peut observer un grand nombre de personnes « au sol »: suite à la chute de tout ou partie des personnes présentes sur le site (quelle qu'en soit la cause : des individus écrasés par la panique, victimes d'un gaz, de la température, d'une explosion, d'un tremblement de terre…). Immédiatement après (niveau B), il faudrait détecter les **précurseurs** de ces accidents de masse, ou les situations dans lesquelles quelques personnes seraient touchées (ex: une voiture roulant vite dans la foule), etc. A un niveau moindre (C), il faudrait détecter et signaler les événements qui pourraient

mettre en danger une ou plusieurs personnes, et finalement les cas générant seulement des pertes d'efficacité sur les flux de personnes (niveau D).

## 2.2 Paradigme logique du système

Dans l'objectif de garantir *que chacun parvienne sain et sauf à sa destination, dans le mouvement collectif*, nous identifions quelques paramètres universels permettant la mesure de l'efficacité de la progression des déplacements:

- ***les Directions de déplacement***, et leur distribution spatiale,
- ***les Vitesses individuelles*** et leur distribution spatiale,
- ***la Densité de la foule***, et sa distribution spatiale.

La mesure de ces paramètres constituera notre **premier** objectif.

Par ailleurs, nous sommes intéressés par un système essentiellement **déterministe**. Un système déterministe est un système bâti sur la causalité. Nous définissons une condition qui implique une règle et nous voulons que cette règle soit toujours activée lorsque cette condition se produit. Ce principe est inhérent à la Loi et à la Sécurité: il n'y a pas de place pour l'aléatoire.

Un système *non-déterministe* est typiquement un système qui serait basé sur une phase "d'entrainement" ou "d'apprentissage" de caractéristiques larges et inconnues ou non-contrôlées de la foule. Par la suite, grâce aux données collectées et organisées à l'intérieur du système, quelques règles apprises déclencheraient une alarme quand une situation observée aurait des caractéristiques similaires à un ou plusieurs cas dangereux appris.

Ce type de système nécessite de larges bases de données d'entraînement et la convergence est difficile à obtenir (surtout lorsque l'objectif de détection est de 100% de précision) : il peut générer de nombreuses fausses alarmes intempestives et peut omettre des évènements importants quand ceux-ci n'ont jamais été appris auparavant. C'est ce qui tend à se produire pour les accidents de foules : ils sont peu communs donc il y a peu de données existantes à ce sujet, ce qui limite les possibilités d'apprentissage.

C'est pourquoi, si une stratégie d'apprentissage paraît intéressante pour améliorer le système sur le long terme, et à mesure que nous identifions de nouveaux signaux faibles, nous proposons qu'un système de supervision des foules soit basé en premier lieu sur des principes déterministes (avec des capacités supervisées d'évolution de l'apprentissage limitées aux signaux faibles, via une supervision manuelle).

A partir de cette méthodologie, nous identifions donc plusieurs règles génériques, en termes de supervision des foules:

- **Détection immédiate d'une "situation dangereuse"** qui peut amener à un avertissement des contrôleurs de foule,
- **Détection immédiate de situations pré-accidentelles, identifiées et connues,** qui ont été établies sur la base de l'expérience, soit du site en surveillance, soit à partir de données plus générales,
- Collecte de statistiques basées sur les paramètres mentionnés antérieurement pour:
  - Prévenir une situation dangereuse (la répartition de la foule en divers points du site peut révéler des risques),
  - Comparer le comportement observé de la foule avec celui d'autres périodes similaires ou d'autres évènements similaires (pour détecter des anomalies, et plus particulièrement **si les dispositions ou les effectifs des contrôleurs de foule ne sont pas appropriés,** ou si les hypothèses réalisées au cours de la *phase de préparation*[3] sont caduques),
  - Comparer les paramètres actuels mentionnés plus haut avec ces accidents appris dans le but de fournir une détection de signaux faibles et aussi d'identifier une situation qui pourrait évoluer vers une situation à risque connue.

## 2.3 Situations dangereuses

Nous avons identifié les situations suivantes (sans doute à compléter avec le retour d'exéprience), par ordre de conséquences décroissant, en suivant les méthodes proposées plus haut (avec le niveau de criticité proposé).

- (A) Détection de chutes massives (de multiples personnes chutent au sol),
- (A) Détection d'un franchissement de seuils mixtes de densité/vitesse susceptibles de provoquer un ou des écrasement(s) affectant un grand nombre de personnes, ou des collisions, ou des chutes dangereuses (quai de gare, escalator)
- (A-B) Détection d'une menace rapide entrant dans la foule (véhicule, …) mettant rapidement en danger plusieurs personnes présentes sur le trajet,
- (B) Détection de fumée provenant de la foule ou atteignant celle-ci (risques incendie / asphyxie),
- (B) Détection d'une ou plusieurs personnes marchant dans une direction non-autorisée (ex : la remontée vers un avion par le couloir de descente),
- (B-C) Détection de dispersion subite d'un groupe dense immobile en plusieurs groupes mobiles (ex : possibilité d'une victime abandonnée ou une manifestation de peur soudaine...),
- (C-D) Détection d'une personne se déplaçant de façon très atypique par rapport au flux d'une foule dense (ex : en travers, ou rapidement au lieu d'attendre son tour) : risque de perturbation, voire risque terroriste,

[3] Cf. Les 4 phases proposées dans le §1.

- (D) Détection d'une personne s'arrêtant ou marchant à contresens de la foule (problème de perturbation dans la progression collective),
- (D) Détection d'un groupe dense et immobile se formant à partir d'une foule mobile et dispersée (impact sur la mobilité collective et risque d'ordre public).

### 2.4 Evènements observés à partir du comportement de la foule

Nous nous proposons également de détecter des situations dans lesquelles la foule observée devient un détecteur d'évènements non visibles du système d'observation (hors du champ des caméras, ou cachés par les premiers rangs). Nous nous proposons donc de détecter des situations telles que:

- (A-B) Un grand nombre de personnes se mettent soudain à courir (fuite) ou s'arrêtent (peur),
- (B-C) Les flux de foule changent brusquement de direction ou de vitesse,
- (C) Détection de « trou » formé dans la foule mobile, les flux de personnes passant de part et d'autre de ce trou (une chute au sol ou un incident peut ne pas apparaitre dans le champ de la caméra à cause de la présence de personnes au premier plan),

De même, (C) la détection de la formation d'un groupe immobile dans une foule mobile, peut révéler une chute au sol, invisible au centre de celle-ci, ou une altercation.

### 2.5 Signaux faibles

Enfin, la recherche de successions d'événements bénins susceptibles d'aboutir à un incident nécessite la mise en œuvre d'un système statistique pour collecter et comparer les situations. Ces successions peuvent être apprises du site lui-même (ex : dans le cadre de la phase de "*Feedback*" mentionné dans le §1, pour améliorer la signalisation) , ou d'une typologie de cas plus large issue de l'expérience collective.

Dans un tel système, des algorithmes d'apprentissage supervisés peuvent aider à collecter des similarités notables qui précèdent généralement des accidents (candidats à devenir des signaux faibles).

## 3. Supervision de la foule par analyse intelligente d'images

Dans le projet CrowdChecker, nous avons développé un démonstrateur de système de supervision de la foule par analyse intelligente d'images issues des caméras de vidéo-surveillance présentes en extérieur ou en intérieur sur les sites où une foule se déploie.

### 3.1 Conditions d'observation de la foule

Nous avons choisi les conditions les plus semblables possibles aux conditions réelles d'un site de transport, de villes, de musées, stades, etc… et utilisé des enregistrements ou des flux de caméras de vidéo-protection couleur déjà installées. Nous avons cherché à réaliser des traitements en temps réel. Nous avons également supposé que la foule est constituée de têtes nues ou dotées de casquettes ou petits chapeaux (pas de masses de grands chapeaux mexicains ou de parapluies), comme on l'observe en général en intérieur, ou en extérieur par une météo correcte (*on notera que dans des conditions climatiques difficiles, par grande chaleur ou forte pluie, la densité de la foule tend à se réduire d'elle même…*).

La majorité de ces caméras sont installées en hauteur, en général entre 2 et 5 m de haut, mais parfois plus, et observent la foule en surplomb.

Nous nous sommes alors intéressés au recouvrement perçu par la caméra entre les individus de cette foule (les premiers cachant partiellement les seconds, etc.).

Supposons qu'une caméra observe une foule en surplomb, celle-ci étant placée sur une zone globalement plane (pas forcément plate). Soit α l'angle de l'axe de la caméra avec la zone plane sur laquelle se tient la foule (en général, dans les installations existantes, on a, $10° < \alpha < 45°$).

Soient alors deux personnes consécutives dans l'axe de la caméra, se suivant à une distance *D* [4]. Si le premier des deux (vers la caméra) est plus grand que l'autre d'une hauteur supérieure ou égale à *D*tan*(α) alors il masquera entièrement l'autre à l'image (voir schéma).

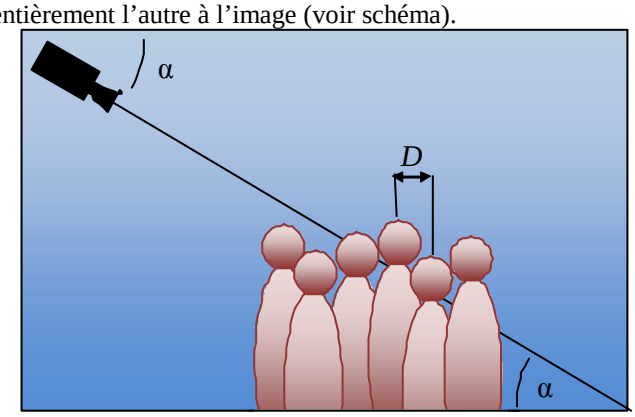



FIG. 1 : Observation de la foule

Supposons maintenant que la taille des personnes dans la foule est répartie selon une loi de Gauss avec une moyenne *m* et un écart-type *s*. La distribution de la différence de taille entre deux voisins suit alors une loi de Gauss de moyenne *0* et variation standard $s* \sqrt{2}$. La probabilité qu'une tête soit totalement masquée par la tête située juste avant dans le champ de la caméra s'obtient par l'expression $D * \tan(\alpha) / (s * \sqrt{2})$ dans une loi de Gauss standard *(0, 1)*. Selon les valeurs de *D* et α, les valeurs probables sont données par les tables de Gauss.

Si nous supposons que l'écart-type dans une population humaine est de *15 cm* (ce qui parait raisonnable pour les hommes, femmes et enfants en occident), nous obtenons les probabilités suivantes de têtes complètement masquées, à

[4] *D* est la distance entre la partie supérieure de deux têtes successives, généralement 50 cm dans une foule dense, 1mètre ou plus dans une foule clairsemée.

mi-hauteur de l'image de la caméra (au niveau où l'angle de vue est α) :

TAB. 1: probabilité de masquage de têtes dans la foule

| Angle de vue moyen | Densité forte (*D=50 cm*) | Densité moyenne (*D=1 m*) |
|---|---|---|
| α=45° | 1% | - |
| α=30° | 8,7% | 0,33% |
| α=20° | 19,5% | 4,4% |
| α=10° | 33% | 20,4% |

Si la caméra est suffisamment loin de la foule (si on peut voir un grand nombre de personnes) cette probabilité peut être interprétée comme la proportion moyenne de têtes masquées à mi-hauteur de l'image.

Ainsi, si nous prenons le cas d'une caméra observant une foule de densité moyenne (*D=1m*), dont l'axe fait un angle de 20° par rapport au sol, et qui a un champ de vision verticale de 20° (et donc 30° d'incidence au sol au bas de l'image, et 10° en haut de l'image), il n'y aura presque aucune tête masquée au pied de l'image, alors que 4,4% des têtes seront masquées à mi-hauteur de l'image (à 20°), et que plus de 20% des têtes seront totalement masquées en haut de l'image (à 10°).

D'autre part, si nous spécifions que pour détecter une tête, nous avons besoin, par exemple, d'une visibilité d'au moins 5 cm[5] du haut de cette tête, alors ces pourcentages augmentent de façon importante (ex : pour *α=20°*, et *D=50 cm*, **27%** des têtes ne sont pas visibles sur au moins 5 cm de hauteur (pour 19,5% des têtes totalement masquées), et **7%** (au lieu de 4,4%) dans le cas d'une foule moyenne si *α=20°* et *D=1 m*).

En conclusion, on notera que les meilleures observations, permettant de bien voir de 90 à 100% des têtes (pour le comptage, la détections d'évènement…) sur des caméras de vidéo-surveillance existantes seront réalisées dans la partie inférieure d'une image prise par une caméra (c'est-à-dire dans la partie de l'image où les angles de vue par rapport au plan du sol sont entre 30° et 90°). Par contre, la détection devient incertaine puis impossible à des angles évoluant entre 20° et 0° par rapport au sol.

Par ailleurs, les têtes des personnes proches de la caméra, au bas de l'image (hors des cas de chapeaux larges ou de parapluies), sont plus faciles à observer (elles sont grandes en pixels, et dans un bon angle de vue, avec une bonne visibilité), tandis que l'observation devient plus difficile lorsqu'elles s'éloignent de la caméra (elles deviennent plus petites en pixels, l'angle de vue plus horizontal entraîne plus de masquages, et la visibilité se dégrade du fait de la distance).

En tant que règle générale, on retiendra que le comptage dans une foule dense observée par une caméra entraîne une sous-estimation de l'ordre de 10 à 40% du nombre de personnes lorsque l'angle au sol varie de 30° à 10° par rapport au sol.

Pour intégrer cette incertitude dans notre démonstrateur, nous avons décidé de mettre en œuvre un modèle de calibration en 3D de la scène observée par l'image de la caméra qui nous permette de prédire la taille en pixels d'une tête à une hauteur donnée de cette image.

## 3.2 Soustraction de fond

De nombreux paradigmes d'analyse intelligente d'images s'appuient sur des techniques de soustraction du fond (ou arrière plan) devant lequel passeraient les personnes, pour les détecter [20]. Ces techniques consistent à établir un modèle de l'arrière-plan « stable » de la scène (par exemple le sol, les murs, le décor...), devant lequel les « mobiles » seraient en mouvement. L'hypothèse sous-jacente utilisée est que la fréquence et la densité des mobiles sont inférieures à l'« observabilité » de l'arrière-plan (la valeur la plus fréquente d'un pixel donné de l'image est celle qu'il a dans l'arrière plan). Le modèle de fond s'actualise périodiquement pour gérer/refléter les changements de lumière (lumière du jour, ombres, jour/nuit...).

Lorsqu'un changement durable survient (ex: quand une voiture se gare sur une place de parking), ce changement est intégré dans l'arrière plan (ou dans une hypothèse à moyen terme de celui-ci), ce qui permet la détection de nouvelles cibles lorsqu'elles passent ensuite devant ce nouvel état de l'arrière plan.

Cependant, on observe que ces hypothèses ne sont pas valables en présence de longues situations de foules denses à l'image. Selon les critères d'un modèle de fond, les foules denses représentent un mouvement perpétuel similaire à un grand objet grouillant devant un fond rarement visible (de plus, les pixels de ce fond sont altérés par l'ombre des passants). L'actualisation du modèle de fond n'a plus de sens à long terme à cause de cette problématique. De plus, le grouillement de la foule en mouvement dans l'image occupe une position globale fixe qui n'apporte pas d'information sur le mouvement et la position des personnes dans ce grouillement (dans un modèle de fond, on espère détecter chaque mobile de façon isolée, avec du fond autour).

En conclusion, nous n'avons pas retenu ces techniques dont les capacités sont insuffisantes pour la gestion de la foule dense.

## 3.3 Placer une croix sur chaque tête

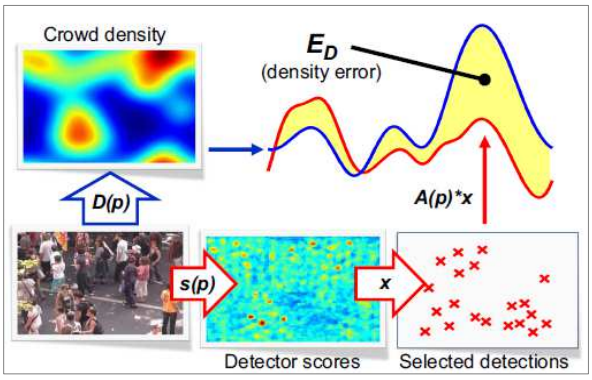
Dans une seconde approche, nous avons cherché à repérer et suivre correctement chaque individu dans une foule, ce qui serait une très bonne réponse à toutes les questions que nous nous posons (cf. §2.2) sur la densité, la vitesse, la direction des passants... Connaître la localisation de chaque personne serait la solution pour détecter presque toutes les situations spécifiées dans le §2.

[5] Ceci dépendra de l'algorithme de détection. Les têtes invisibles ne peuvent être comptées mais il y a aussi une taille minimum de tête qui doit être présente à l'image pour permettre la détection. Nous verrons ensuite que, par exemple, pour détecter la forme Ω tête-épaules, 25 à 35 cm seront requis au lieu des 5 cm supposés ici dans ce cas.

Plusieurs approches ont été étudiées afin d'identifier et suivre différentes combinaisons de parties du corps (tête, épaule, poitrine, bras…) qui font partie de la foule [3, 4, 5, 6, 7, 8, 9,10]. Certaines de celles-ci ont abordé la recherche de formes ellipsoïdales, d'autres ont abordé l'apprentissage de formes de têtes, la forme d'une tête sur les épaules d'un individu ressemblant au caractère grec Ω. C'est pourquoi de nombreux travaux étudient la recherche spécifique de la forme Ω [4].

Dans le projet CrowdChecker, l'équipe chargée du projet a tenté de reproduire et d'améliorer ces algorithmes comme une approche générale pour la détection et le suivi des têtes [11]. Nous avons cherché à mesurer leurs performances et leurs capacités à fonctionner en temps réel sur une architecture de type ordinateur PC, qui pourrait constituer une plate-forme raisonnable en termes de coûts[6] pour un outil de supervision.

Dans [11], l'utilisation d'une estimation de la densité des personnes basée sur une régression [24] a été étudiée afin d'améliorer la détection de têtes et le suivi de personnes sur une scène de foule. Dans cette approche, une carte de densité a été utilisée pour prévoir la localisation des têtes.



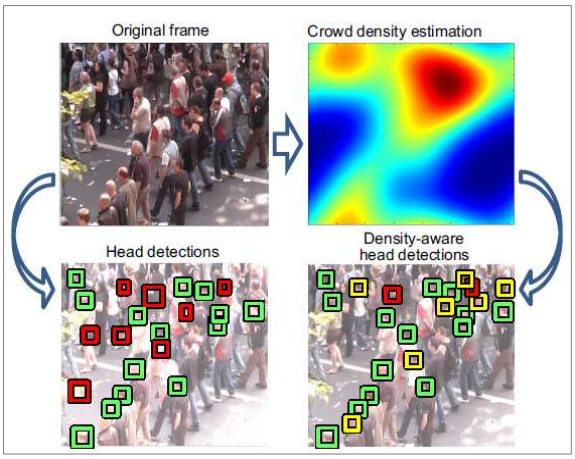




FIG. 2: vision générale d'un modèle de détection de personnes

La détection de personnes est vue comme une minimisation d'énergie (Eq. 3.1), qui implique une fonction d'énergie combinant les résultats : des détections d'individus $E_S$, des restrictions de non-superposition $E_P$, et des restrictions imposées par la densité estimée de personnes dans une scène $E_D$ (cf. Figure 2):

$$\underset{x\in\{0,1\}^N}{\operatorname{argmin}} \quad (\overbrace{-s^T x}^{E_S} + \overbrace{x^T W x}^{E_P} + \overbrace{\alpha||D-Ax||_2^2}^{E_D})$$

(Equation 3.1)

ou $x$ est un vecteur à *N* dimensions représentant toutes les détections sur l'image entière ($x_i = 1$ si la détection sur une localisation $p_i$ est valide), $s$ est un vecteur à N dimensions représentant un indicateur de confiance pour chaque localisation $p_i$, $W$ est une matrice $N \times N$ où $W_{i,j} = \infty$ si les détections sur les localisations $p_i$ et $p_j$ se recouvrent significativement, et $W_{i,j} = 0$ sinon, $D$ est un estimateur de densité basé sur la régression, et $Ax$ est un produit matriciel représentant une évaluation de la densité des détections actives. La matrice $N \times N$ $A$ est conçue de façon que chaque colonne $A_i$ est une fenêtre gaussienne de taille $\sigma$ et centrée en $p_i$.

Plus simplement, l'équation. 3.1 vise à détecter plus de têtes là où la foule est dense, et à éviter de fausses détections quand la foule est clairsemée.

Nous avons obtenu des progrès significatifs de performances en détection de têtes et en tracking sur des vidéos de foule complexes présentant une densité hétérogène (voir Fig. 3), par rapport notamment au détecteur de têtes de référence [23] (*courbe a*), ou à ce dernier combiné avec des filtres géométriques prenant en compte la taille des têtes, ou enfin au détecteur de base combiné avec des contraintes de cohérence temporelles du mouvement des têtes (*courbes b, c*). Le détecteur prenant en compte la densité (*courbe rouge d*) dépasse sans aucun doute les trois autres détecteurs. Le détecteur utilisant

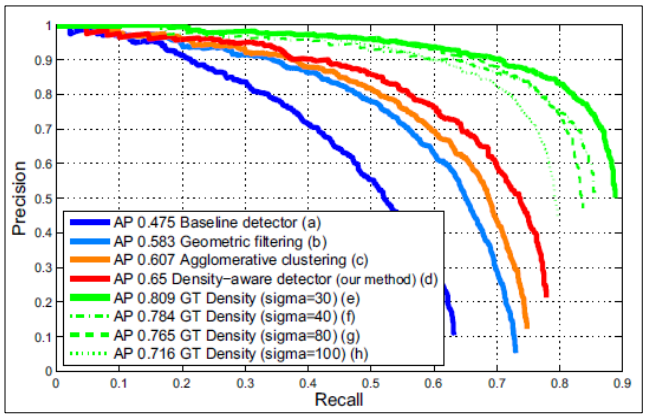



FIG. 3: Courbes de précision / rappel

l'*estimation de densité issues de la vérité terrain* obtenu par lissage gaussien de la distribution des personnes issue de la vérité terrain, a été aussi représenté (courbes vertes e-h) et révèle les bénéfices du détecteur prenant en compte la densité quand il est utilisé conjointement avec une méthode

[6] Le coût d'un PC performant est de l'ordre de 1000 à 2000 €. Analyser 4 à 8 caméras sur celui-ci représente donc un coût « matériel » de 250 à 500 €, auquel il faut ajouter le cout de la licence du logiciel, et de l'installation et la configuration. Ceci est à comparer avec le coût brut d'une caméra (100-1000 €), et à celui d'une caméra posée et raccordée (~1 à 10 K€ selon le site).

d'estimation de densité performante (ceci n'étant toutefois qu'une technique de validation).

Malgré de nombreuses améliorations proposées par l'équipe de recherche du projet, il a été démontré qu'il n'était pas possible de détecter plus de 50% des têtes moyennant la détection de 15% de fausses têtes, ou bien la détection de 70% des têtes moyennant la détection de 40% de fausses têtes (*courbe rouge*). De plus, l'obtention de ces résultats peu satisfaisants nécessite un temps de calcul très élevé avec un matériel puissant.

Cependant, cette approche peut être intéressante quand on l'utilise sur une zone de détection réduite. Nous en avons retenu le principe, pour une utilisation locale, avec les limites de visibilité des têtes indiquées au §3.1. Il faut signaler que pour observer une forme complète Ω sur une tête, il est nécessaire de voir environ 30 cm du haut du corps, ce qui réduit de manière significative la proportion d'individus détectables, notamment au fond de la scène.

## 3.4 Apprentissage de patchs de situation de foule

Au cours du projet CrowdChecker, l'équipe de développement a également étudié une approche totalement différente, orientée vers l'apprentissage d'une multitude de petits patches extraits de multiples vidéos de foule (petites séquences vidéos réduites à une portion de l'image d'origine, et sur une courte durée). Dans cette approche [12], ces patchs ont été classés selon une stratégie d'apprentissage, afin de constituer un noyau permettant de « reconnaître » les éléments constitutifs d'une vidéo quelconque de la foule.

Cette approche nous a semblé assez utile pour prévoir le parcours d'un individu dans une foule, mais peu fiable dans le cas de la foule dans son ensemble. De plus, elle était extrêmement consommatrice de ressources matérielles.

## 3.5 Analyse du mouvement

Finalement, nous avons étudié plusieurs approches d'estimation du mouvement [13, 14, 19], avec pour objectif l'identification de zones de foules en mouvement cohérent (vitesse et direction). Nous avons réalisé une estimation de ce mouvement à différentes échelles de temps, à cause du mouvement « Brownien » des personnes et des membres (bras, jambes) dans une foule.

Bien qu'il ait été démontré [15] que des règles très simples gouvernent les mouvements des personnes dans la foule, une foule n'est pas constituée de sous-groupes homogènes où chaque individu présente les mêmes gestes (sauf lors d'un défilé militaire…).

Différentes méthodes d'estimation de mouvement ont donc été évaluées pour cette tâche:

- Méthodes de type flot optique: méthode différentielle basée sur l'hypothèse que la luminosité (*I*) d'un point mobile (*x*, *y*) est constante au cours du temps *t* :

$$I(x, y, t) = I(x + \Delta_x, y + \Delta_y, t + \Delta_t) \quad .$$

  Et en complément :
  - Horn et Shunck [26] supposent la régularité du flot global,
  - Lucas et Kanade [27] supposent que le déplacement d'un pixel p est faible et lié à celui des pixels voisins (dans une fenêtre prédéterminée). De plus, un algorithme de détection de points saillants de Shi et Tomasi [28, 29] peut être appliqué avant l'estimation du mouvement afin de sélectionner les points d'intérêt sur lesquels l'estimation de mouvement sera appliquée. Cette méthode est très populaire pour l'analyse des flux de foule [21].
- Le réagencement par blocs (Block matching) consiste à diviser l'image en petits blocs se recouvrant partiellement et à les apparier entre images successives selon un critère de similarité (en général, la somme des écarts absolus). [22] présente une méthode générale de détection et suivi de la foule s'appuyant notamment sur cette technique.

Lorsqu'on compare ces méthodes, un inconvénient majeur de la méthode de Horn et Shunck est son incapacité à extraire des mouvements singuliers, du fait de son hypothèse de régularité générale du flux. Ce n'est pas le cas des deux autres méthodes mentionnées ici, malgré le fait que celles-ci provoquent des fausses détections au sein de régions homogènes.

Dans l'absolu, nous souhaitions détecter les flux principaux de la foule (vert), tout en préservant la détection le longs parcours individuels qui ne suivraient pas ces flux.

Par conséquent, nous avons appliqué une intégration des mouvements individuels pour identifier les principaux flux de la foule, et détecté les singularités plus petites à l'aide d'un complément de tracking.

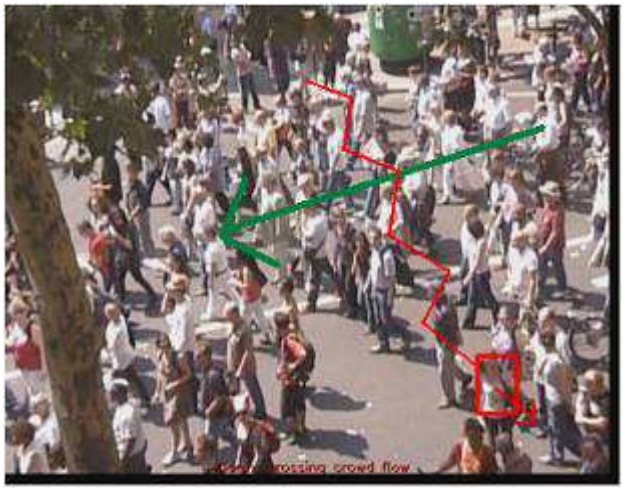


FIG. 4: Suivi de mouvement de foule

Cette approche s'est montré coûteuse en temps de calcul mais accessible en temps réel. Certaines propriétés issues

de l'encodage des flux d'images numériques compressés (ex : H.264…) peuvent aider à réduire ce cout.

Enfin, lorsque les parties de la foule s'immobilisent, la méthode doit être complétée par la détection de personnes immobiles. On doit pour cela différencier un groupe immobile au sein d'une foule mobile, d'un trou dans la foule. Pour cette raison, un **modèle de tracking de long terme** a été imaginé pour capitaliser des informations sur les personnes immobiles.

## 3.6 Estimations de densité

Ensuite nous avons complété cette approche en mettant au point un modèle de **densité**, afin d'avoir une appréciation précise du nombre de personnes présentes, de leur direction et vitesse. Nous avons étudié plusieurs modèles différents avec comme objectif d'estimer la granularité, comme par exemple [16, 17, 18, 24].

Plus précisément, dans [24] il est proposé un cadre général pour un apprentissage supervisé de l'estimation de densité. On émet l'hypothèse d'une série d'images d'entrainement avec des cartes de caractéristiques denses $\phi(p) \in \mathrm{R}^m$ sur chaque pixel p, et on utilise la vérité terrain des positions des têtes ($\xi$ ). Les fonctions de densité dans cette approche sont des fonctions réelles sur pixels, dont les intégrales sur les régions de l'image correspondent au nombre d'éléments inclus. Pour chaque image d'entrainement I, la fonction de densité vérité terrain se définit comme une fonction d'estimation par noyaux basée sur les points suivants:

$$\forall p \in I, \qquad F^0(p) = \sum_{P \in \xi} \mathcal{N}(p; P; \sigma^2 \mathbf{1}_{2\times 2})$$

où $\mathcal{N}(p; P; \sigma^2 \mathbf{1}_{2\times 2})$ est un noyau Gaussien 2D normalisé, évalué en p, dont la moyenne est centrée sur le point P, avec une matrice de covariance non orientée. Etant donné cet ensemble d'images d'entrainement I assorti des fonctions de densité issues de la vérité terrain, la transformation linéaire d'une représentation de caractéristiques qui se rapproche de la fonction de densité à chaque pixel est apprise :

$$\forall p \in I, \qquad F(p|w) = w^T \phi(p)$$

où $w \in \mathrm{R}^m$ est un vecteur de paramètre de la transformation linéaire obtenue des données d'entrainement k qui minimise la distance de MESA.

Une fois l'apprentissage effectué, une estimation par comptage d'objets peut être obtenue sur chaque pixel ou sur chaque région donnée en intégrant cette fonction sur la surface de la zone d'intérêt.

Le modèle de densité permet alors de développer, pour des images de la foule, des fonctions de comptage du flux de personnes traversant une ligne tout comme des fonctions de comptage du nombre de présents dans une zone.

## 3.7 Résultats du projet

Nous avons réalisé un démonstrateur qui a été testé sur des vidéos enregistrées et sur des flux en temps réel. De nombreuses études antérieures sur l'analyse de la foule ont abordé des cas de foule clairsemée (Figure 5), et nous voulions aussi aborder des situations de foules denses et mobiles. Nous avons testé nos algorithmes sur les deux types de vidéos. Nous avons produit plusieurs vidéos avec leur vérité terrain [11, 12] qui sont disponibles pour de futures recherches. Notre système estime avec exactitude la densité de la foule des situations de densité moyenne 2.1 personnes/m$^2$ (soit 1.5 à 3 fois plus denses que les vidéos généralement utilisées dans la littérature pour ces méthodes). On trouvera dans certaines situations réelles des densités supérieures (jusqu'à 7 à 9 par m$^2$), mais alors le mouvement est considérablement lent, typiquement de quelques mètres par minute: les contacts entre les personnes limitent leur capacité de mouvement [30].

FIG. 5: la foule dans la littérature

De plus, le démonstrateur réalisé détecte automatiquement en temps réel les contresens, sans hypothèse particulière sur la cible (même partiellement occluse ou couvrant partiellement la foule traversée). Cette détection s'applique à des cibles à partir de 10-20 pixels de côté (la taille dépendant du contraste). Le démonstrateur donne de très bon résultats sur les fonctions implémentées telles que le comptage, la détection de contresens, la mesure de vitesse, la détection du franchissement d'un seuil de vitesse, de densité… Quelques exemples d'image de détection sont présentés ici :

(a) Carte de densité

(b) Personne traversant le flux de personne dans une foule

(c) Accélération subite d'une foule


FIG. 6: exemples

Plusieurs fonctions du démonstrateur sont regroupées en Figure 7. Les angles de vue sur l'horizon sont de 43,5° au pied de l'image et de 17,3° au fond en haut de l'image (cf. §3.1). La calibration 3D permet la mesure de la superficie du sol dans la zone dessinée (24.1 $m^2$). Les estimations de comptage et densité dans cette zone sont affichées (1.0 pers/$m^2$), avec comptage des flux de personnes traversant les deux lignes dessinées (à gauche et à droite), et une alarme est levée à propos d'une personne qui traverse le flux (carré bleu). Sur la seconde image, nous observons les vecteurs de direction et les estimations de vitesse des flux principaux de personnes (violet) et de la cible signalée (en vert).

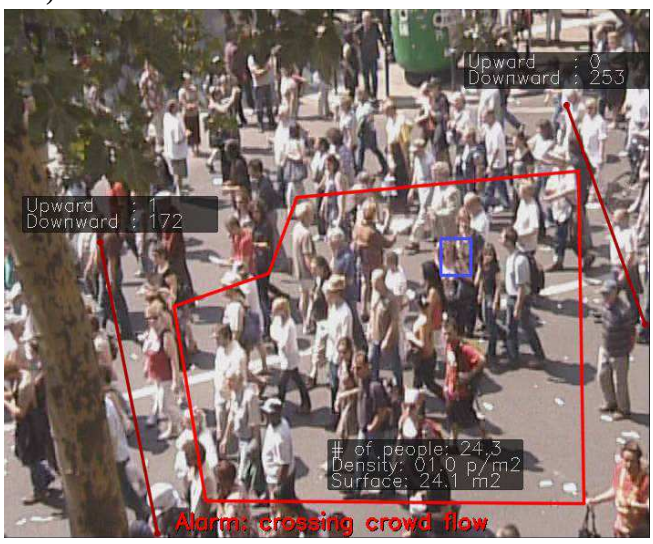



FIG. 7: (En haut) mesure simultanée de densité, comptage et détection. (A gauche) carte de vitesses et directions.

La vérité terrain sur la densité dans la zone dessinée est de 1.04 pers/$m^2$ (l'erreur est de 4%). La vitesse instantanée du flux principal de personnes est mesurée à 2.1 km/h, alors que la vitesse vérité terrain à long terme est d'environ 2.24 km/h (erreur = 6.6%). Nous étudions l'utilisation de filtre particulaire pour améliorer celle-ci.

Nous travaillons à améliorer le démonstrateur et à lui apporter de nouvelles fonctions comme par exemple la détection de la fumée dans la foule, d'un véhicule entrant dans cette même foule et d'autres applications présentées dans les §2.3 et §2.4.

# 4. Perspectives

La détection de dangers dans les foules est un objectif très ambitieux.

Comme nous l'avons vu précédemment, les besoins fonctionnels de détection présentés dans cet article ne sont pas encore inclus dans les appels d'offres. Il sera probablement assez long de les y voir apparaître.

De plus, à cause à la nature intrinsèque de la foule, chaque situation de foule possède une typologie conséquente d'accidents potentiels. Imaginer tous ces risques est relativement angoissant. Un outil de supervision intelligent tel que nous l'avons présenté fait évoluer la responsabilité des agents de surveillance de la foule ainsi que celle de leurs organisations : ils ne pourront plus indiquer qu'ils n'ont pas vu l'accident si le système les en a prévenus. Leur responsabilité pour assister en cas de danger est augmentée, notamment dans les cas où une alarme reçue n'aurait pas été gérée.

Il se pourrait que des considérations d'ergonomie des outils, et de charge de travail soient soulevées à l'usage, comme pour les pilotes de ligne.

Il est probable qu'une fois passées les réticences à l'usage de ce type d'outil, ces outils amélioreront en premier lieu la phase de *Feedback* (cf. §1) de la gestion pacifique des foules, puis la phase de *Préparation*, et plus tard, aideront l'évolution de l'organisation et des procédures des opérateurs de supervision des foules, pour améliorer la phase *d'Actions*.

Nous sommes optimistes sur le fait que la sécurité globale en sera améliorée.

# 5. A propos

Nous souhaitons remercier ici Jacques Blanc-Talon et Veronique Serfaty de la DGA qui ont soutenu notre travail, ainsi que les contributeurs de la SNCF et de la Préfecture de Police de Paris, pour leur aide aux spécifications. Nous souhaitons également remercier l'équipe de recherche Willow pour sa participation et son travail de recherche, et Y-O Renault pour son apport en maths.

L'équipe française de recherche Willow est une équipe mixte regroupant CNRS, INRIA, et ENS, constituée d'environ 40 personnes et dirigée par Jean Ponce.

EVITECH est une PME, membre du pôle de compétitivité **SYSTEM@TIC Paris Région**. Elle étudie et développe des systèmes innovant de traitement d'image pour la sécurité globale.

P. Bernas est Ingénieur Civil des Mines et Docteur en informatique. P. Drabczuk est diplômé de l'Institut d'Optique Graduate school. G. Née termine une thèse en traitement d'image avec le GREYC / Université de Caen.

# Bibliographie

[1] European communities, *Meeting the challenge, the European Security Research Agenda, a report from the security research advisory board*, sept. 2006.

[2] RTCA DO-178B, *Software Considerations in Airborne Systems and Equipment Certification*, RTCA Inc., Washington D.C, 1992 / ED 12B, EUROCAE, Paris, 1992.

[3] Ben Benfold, Ian Reid, Stable, *Multi-Target Tracking in Real-Time Surveillance Video*, CVPR 2011

[4] Tao Zhao, Ram Nevatia, *Stochastic Human Segmentation from a Static Camera*, CVPR 2004

[5] Junliang Xing, Haizhou Ai and Shihong Lao , *Multi-Object Tracking through Occlusions by Local Tracklets Filtering and Global Tracklets Association with Detection Responses* , CVPR 2009

[6] Chang Huang, Bo Wu, and Ramakant Nevatia , *Robust Object Tracking by Hierarchical Association of Detection Responses* , ECCV 2008

[7] Vivek Kumar Singh, Bo Wu, Ramakant Nevatia , *Pedestrian Tracking by Associating Tracklets using Detection Residuals* , WMVC 2008

[8] Bo Wu and Ram Nevatia , *Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian* , IJCV 2007

[9] Bo Wu, Ram Nevatia and Yuan Li, *Segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses*, CVPR 2008

[10] Zhao T. and Nevatia R., *Tracking multiple humans in crowded environment*, CVPR 2004

[11] M. Rodriguez, I. Laptev, J. Sivic and J.Y. Audibert, *Density-aware person detection and tracking in crowds*, ICCV 2011

[12] M. Rodriguez, J. Sivic, I. Laptev and J.Y. Audibert, *Data driven crowd analysis in videos*, ICCV 2011

[13] Saad Ali and Mubarak Shah, *A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis*, CVPR 2007

[14] Min Hu, Saad Ali and Mubarak Shah , *Learning Motion Patterns in Crowded Scenes Using Motion Flow Field*, International conference on pattern recognition 2008

[15] Mehdi Moussaıd, Niriaska Perozo, Simon Garnier, Dirk Helbing, Guy Theraulaz, *The Walking Behaviour of Pedestrian Social Groups and Its Impact on Crowd Dynamics*, PLoS ONE 2010

[16] Siu-Yeung Cho, T.W.S. Chow and Chi-Tat Leung, *A neural-based crowd estimation by hybrid global learning algorithm*, IEEE Transactions on Systems, man, and cybernetics 2002

[17] A.B. Chan, Z.S.J. Liang and N. Vasconcelos, *Privacy preserving crowd monitoring: counting people without people models or tracking*, CVPR 2008

[18] D. Ryan, S. Denman, C. Fookes and S. Shridharan, *Crowd counting using multiple local features*, Digital image computing 2009

[19] Ovgu Ozturk, Toshihiko Yamasaki, Kiyoharu Aizawa, *Detecting Dominant Motion Flows In Unstructured/structured Crowd Scenes*, , ICPR 2010

[20] T. Bouwmans, F. E. Baf, and B. Vachon. *Statistical background modeling for foreground detection: A survey*. Handbook of Pattern Recognition and Computer Vision World Scientific Publishing, 4:181–199, Jan. 2010.

[21] Anil M. Cheriyadat and Richard J. Radke, *Automatically determining dominant motions in crowded scenes by clustering partial feature trajectories*, International Conference on Distributed Smart Cameras 2007

[22] Sergio A. Velastin, Boghos A. Boghossian b, Maria Alicia Vicencio-Silva, *A motion-based image processing system for detecting potentially dangerous situations in underground railway stations*, Transportation Research Part C: Emerging Technologies 2006

[23] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part based models*, IEEE PAMI 2010.

[24] V. Lempitsky and A. Zisserman. *Learning to count objects in images*, NIPS 2010.

[25] WorkSafe Victoria (Australia), *Crowd control at venues and events, a guide to support and assist crowd control agencies*, available at http://www.worksafe.vic.gov.au/

[26] B.K.P. Horn and B.G. Schunck, *Determining optical flow*, Artificial Intelligence 1981

[27] B. D. Lucas and T. Kanade, *An iterative image registration technique with an application to stereo vision*, Imaging Understanding Workshop 1981

[28] J. Shi and C. Tomasi, *Good Features to Track*, CVPR, 1994

[29] C. Tomasi and T. Kanade, *Detection and Tracking of Point Features*, Pattern Recognition 2004

[30] Dirk Helbing, Anders Johansson, HE Habib Z. Al-Abideen, *Crowd turbulence: the physics of crowd disasters*, 5th International conf. on non-linear mechanics (ICNM-V), Shanghai, June 2007.