Chapitre 7

Reconnaissance d'activités/d'actions

Plan du chapitre

- Reconnaissance d'activités/d'actions
  - Par détection et suivi
  - Par historique de mouvement
  - Par points caractéristiques et approche sac de mots visuels
- Détection du transport d'objets

Reconnaissance d'activités

- Par détection et suivi #1
  - Identification d'activités basée sur la trajectoire et la séparation/fusion de blobs.
  - Débute par une soustraction d'arrière-plan basée sur le contraste de la luminance.
  - Le contraste de luminance est la différence relative entre la luminance de l'objet et la luminance de l'arrière-plan.
  - Calculé avec le *Y* de *YUV*:

$$Y = 0.299 * R + 0.587 * G + 0.114 * B$$
$$U = 0.436 * (B - Y) / (1 - 0.114)$$
$$V = 0.615 * (R - Y) / (1 - 0.299)$$

Reconnaissance d'activités

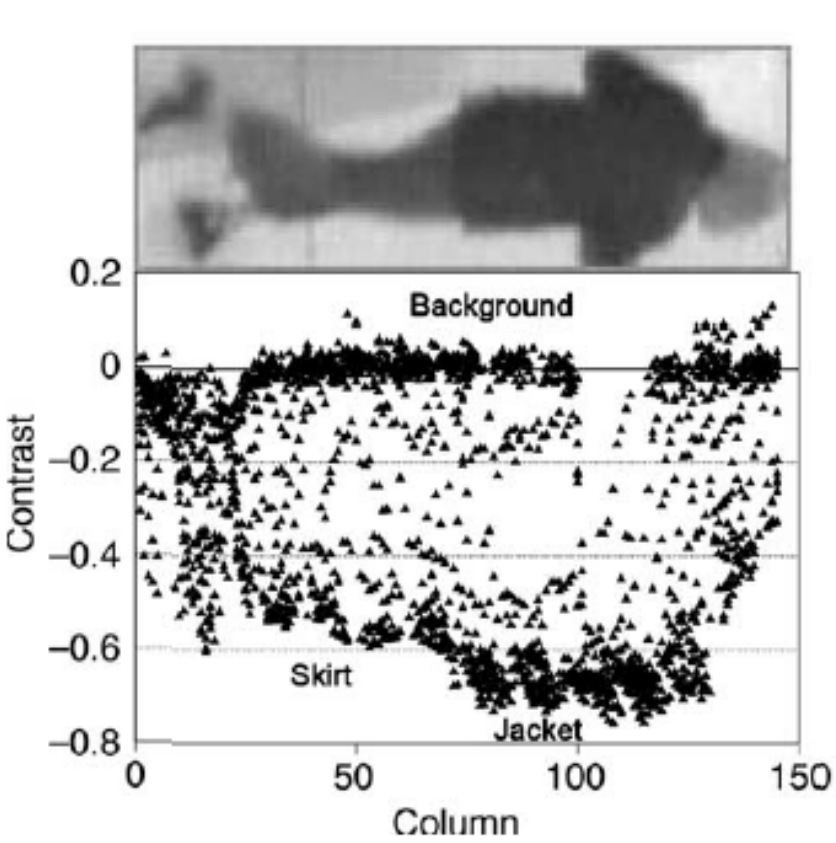
- Par détection et suivi #1 (*suite*)
  - Le contraste de luminance *C(i,j)* est calculé par:

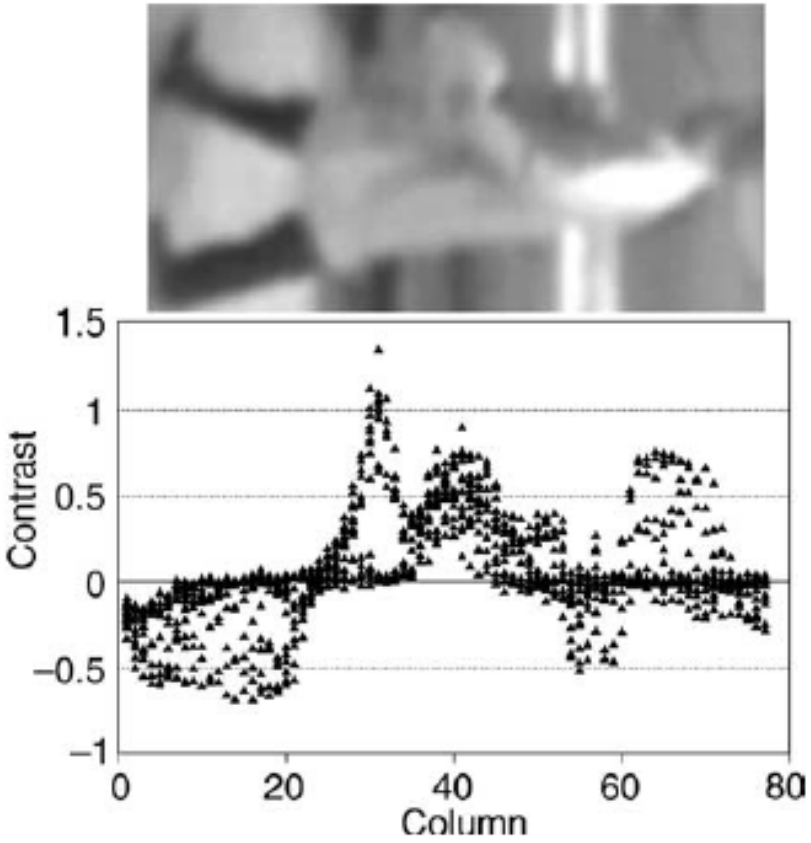
$$C(i,j) = \frac{y(i,j) - y_B(i,j)}{y_B(i,j)}$$

Les valeurs vont de [-1,254], avec des valeurs négatives représentant les objets plus foncés que l'arrière-plan, et les valeurs positives pour l'arrière-plan plus foncé que les objets. Si $y_B(i,j)=0$, on met $y_B(i,j)=1$.

Reconnaissance d'activités

- Par détection et suivi #1 *(suite)*





Reconnaissance d'activités

- Par détection et suivi #1 (*suite*)
  - Pour chaque blob détecté, la trajectoire est calculée à l'aide des intersections des rectangles englobants des blobs de deux images consécutives.
  - La trajectoire tient aussi en compte les séparations/fusions de blobs aussi établies à l'aide des rectangles englobants et d'une matrice d'intersection de blobs.

Reconnaissance d'activités

- Par détection et suivi #1 (*suite*)

$$M_{t-1}^{t} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad M_{t}^{t-1} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$S_{t-1}^{t} = \begin{bmatrix} 1 & 1 & 2 & 4 \\ & & 3 & \end{bmatrix} \quad S_{t}^{t-1} = \begin{bmatrix} 1 & & & \\ 2 & 3 & 3 & 4 \end{bmatrix}$$

Reconnaissance d'activités

- Par détection et suivi #1 (*suite*)
  - Connaissant les fusions/séparations et les trajectoires, on peut faire la détection de certains événements.
  - On suppose qu'il y a peu de personnes (donc moins d'occlusions).
  - La détection des événements est faite par des règles.

Reconnaissance d'activités

- Par détection et suivi #1 (*suite*)
  - Exemple: Bagage abandonné
    - Un blob se sépare en deux;
    - Un des deux blobs est stationnaire, l'autre s'éloigne;
    - On peut définir une période de temps sur la durée de la séparation.
  - Exemple: Une personne se cache
    - Un blob disparaît pendant plusieurs images;
    - La position du dernier centroïde est loin d'une porte;
    - La position du dernier centroïde est près d'un endroit où on peut se cacher.
    - On peut définir une période de temps sur la durée de l'absence de la personne.

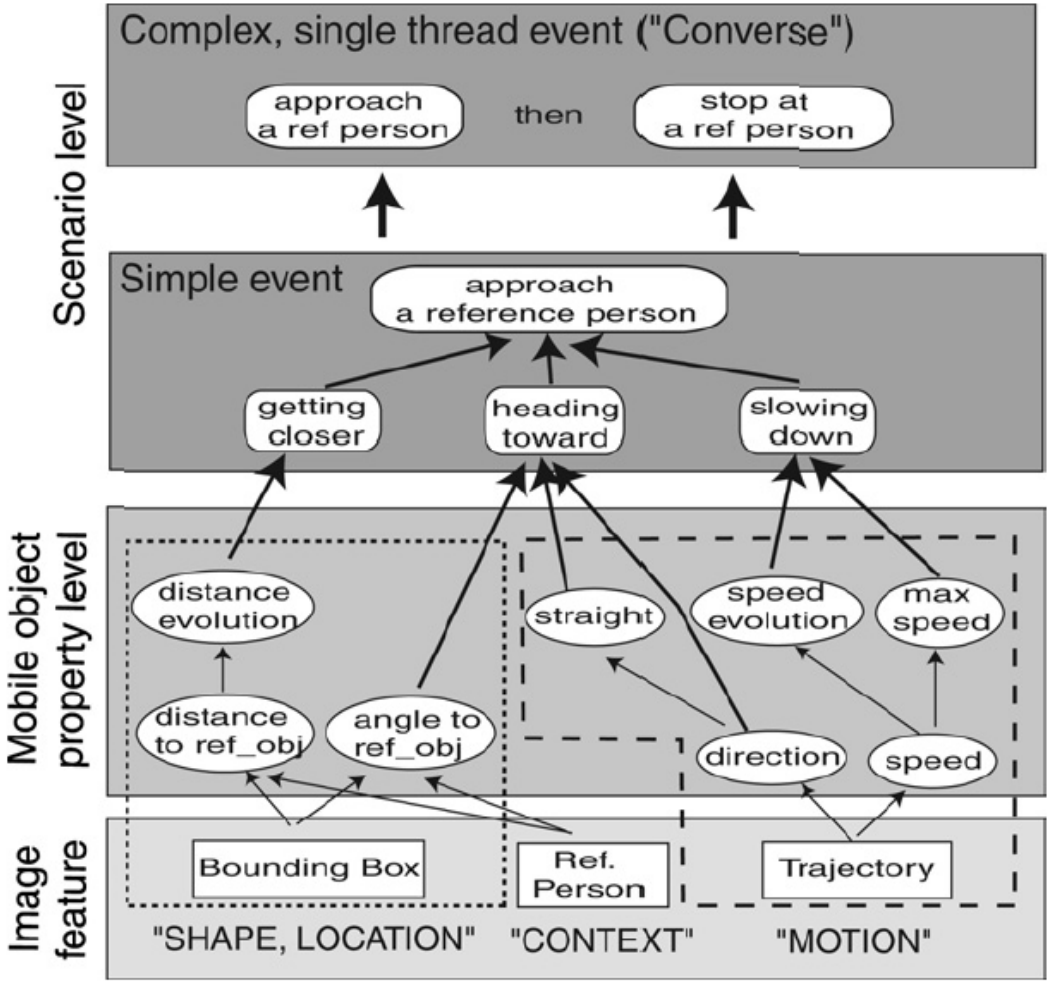
Reconnaissance d'activités

- Par détection et suivi #1 (*suite*)

Reconnaissance d'activités

- Par détection et suivi #2
  - Un événement est exprimé comme une hiérarchie d'actions (ou entités).
  - Il y a deux niveaux de traitement:
    - Bas niveau: Détection des objets et calcule de la trajectoire.
    - Haut niveau: interprétation des trajectoires comme des actions, et ensuite en scénario.

Reconnaissance d'activités

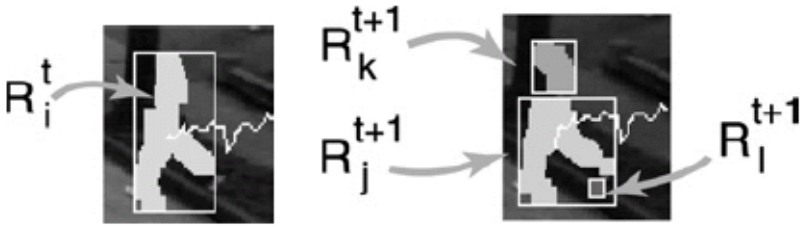


Reconnaissance d'activités

- Par détection et suivi #2 (*suite*)
  - Après la soustraction d'arrière-plan, pour trouver les trajectoires, le suivi est fait avec les distributions d'intensité des objets (par apparence) et la connaissance du niveau du sol.
  - Le niveau du sol fait partie du contexte d'opération des algorithmes. C'est le contexte spatial. Permet de résoudre les séparations et les fusions.

Reconnaissance d'activités

- Par détection et suivi #2 (*suite*)



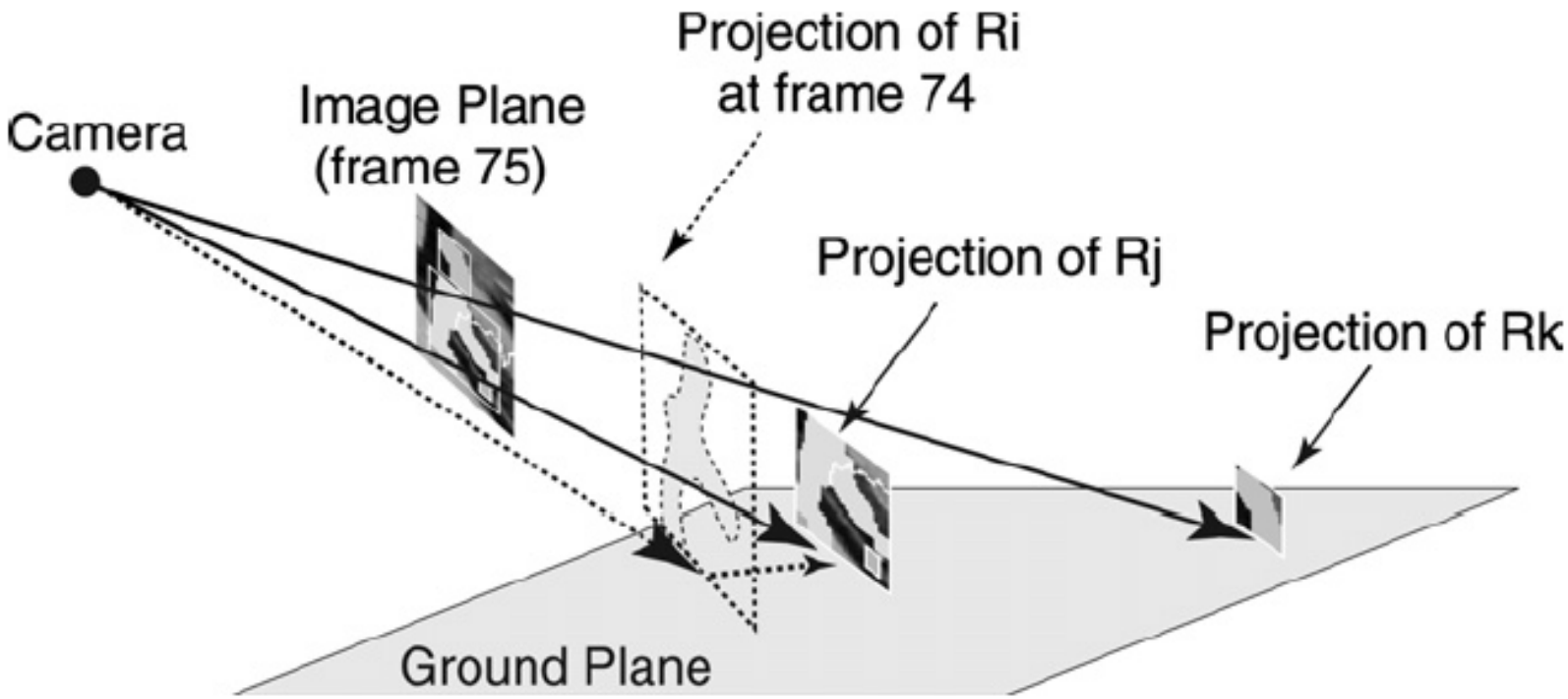
(A) Frame 74. (B) Frame 75.

Fig. 3. Splitting of moving regions and noise.

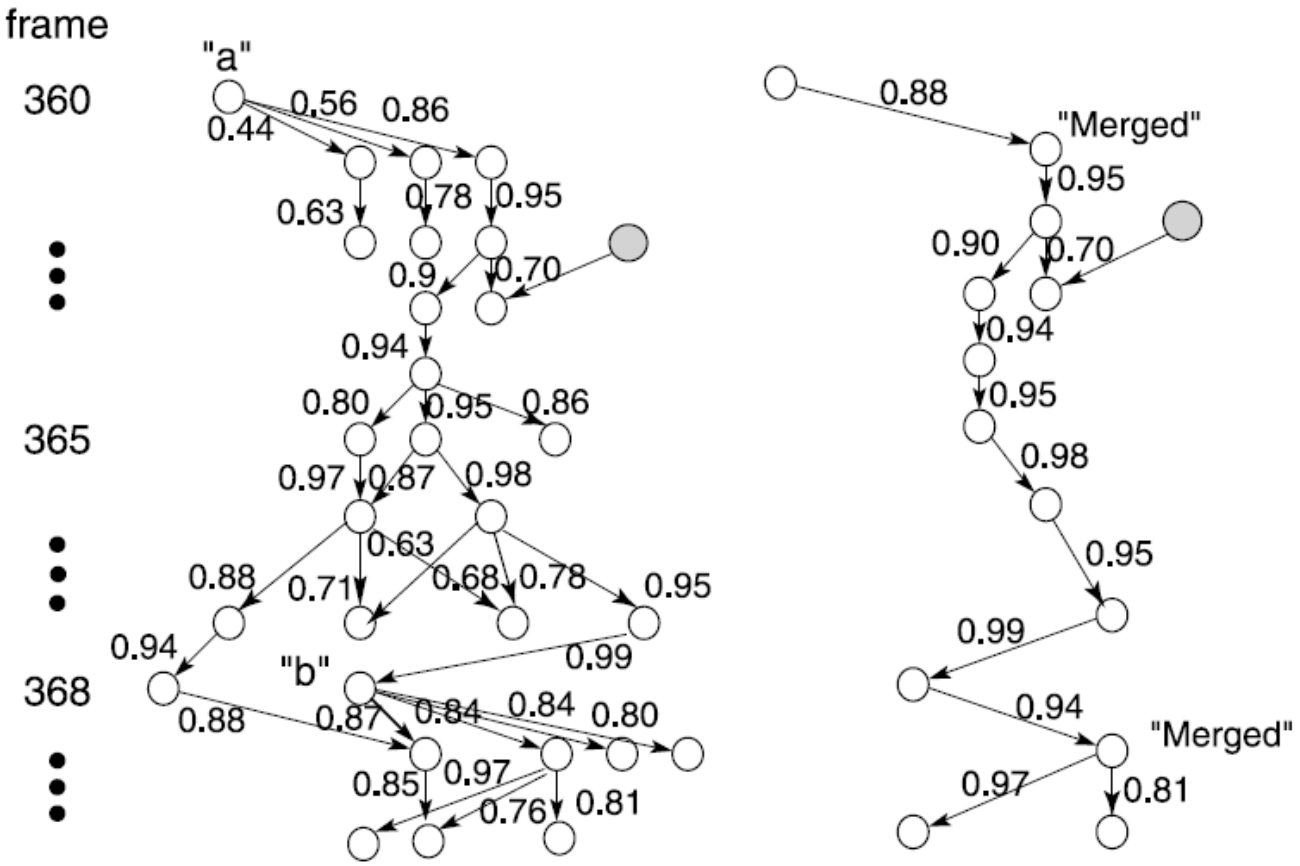
  - Calibration de la caméra pour trouver la matrice d'homographie projective *H*. Et on peut ensuite tester des hypothèses d'objets fragmentés ou fusionnés en les projetant au sol.

Reconnaissance d'activités

- Par détection et suivi #2 (*suite*)



Reconnaissance d'activités



Sans le sol

Avec le sol

Reconnaissance d'activités

- Par détection et suivi #2 (*suite*)
  - Dépendamment du type de scénario, la représentation varie, mais c'est toujours le principe des machines à états.
  - Pour les scénarios plus simples:

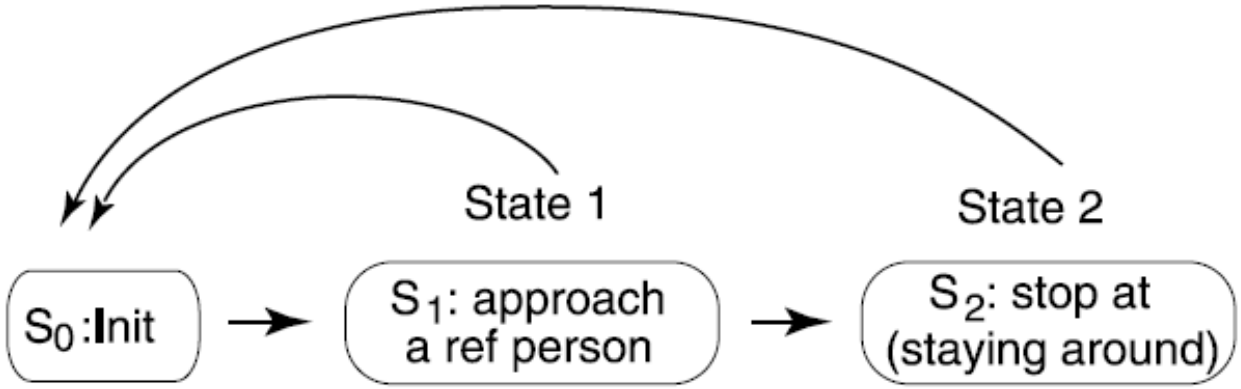


Fig. 9. A finite-state automaton that represents the complex event "converse."

Reconnaissance d'activités

- Par détection et suivi #2 (*suite*)
  - Pour les scénarios plus complexes, un graphe de relation temporelle binaire est utilisé.

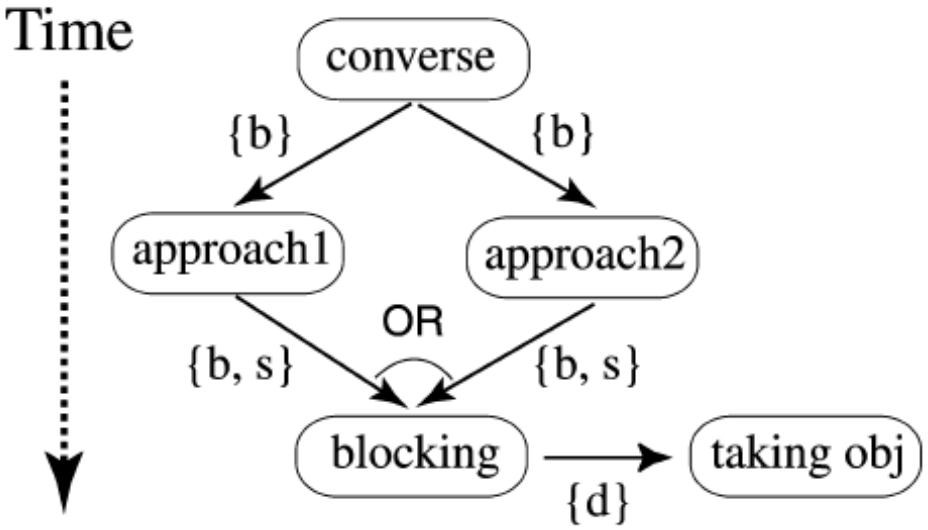


Fig. 10. A graphical description of a multi-thread event.

Reconnaissance d'activités

- Par historique de mouvement
  - Les événements sont caractérisés principalement par le changement local des pixels dans un voisinage sur une période de temps.
  - Une activité est représentée par un groupe corrélé d'événements.
  - Les activités sont apprises par classification.

Reconnaissance d'activités

- Par historique de mouvement (*suite*)
  - Les changements locaux sont modélisés par le modèle PCH (*pixel change history*).
  - On cherche à obtenir la direction du changement, la forme du "changement" et sa localisation. PCH capture tout ça.
  - Il y a trois types de changements significatifs:
    - Court terme: Déplacement
    - Moyen terme: Personne qui s'arrête
    - Long terme: Ajout/retrait d'un objet statique

Reconnaissance d'activités

- Par historique de mouvement (*suite*)
  - PCH:

$$P_{\varsigma,\tau}(x, y, t) = \begin{cases} \min\left(P_{\varsigma,\tau}(x, y, t-1) + \frac{255}{\varsigma}, 255\right) \\ \quad \text{if } D(x, y, t) = 1 \\ \max\left(P_{\varsigma,\tau}(x, y, t-1) - \frac{255}{\tau}, 0\right) \\ \quad \text{otherwise} \end{cases}$$

  - $P_{\zeta,\tau}(x,y,t)$ est le PCH pour un pixel (x,y), $D(x,y,t)$ est une image binaire indiquant l'avant-plan, ς est le facteur d'accumulation, τ est le facteur d'atténuation. Le ratio de ς et τ détermine le poids attribué aux changements récents.

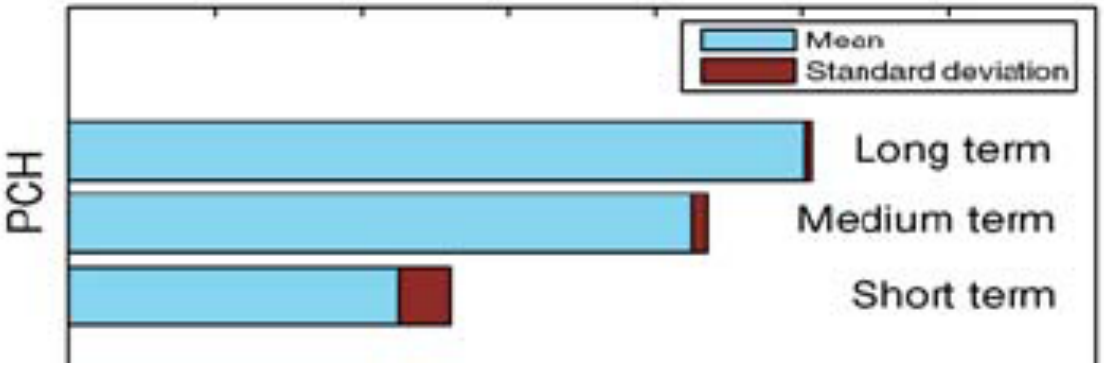
Reconnaissance d'activités

Different types of visual changes

PCH

Reconnaissance d'activités

- Par historique de mouvement (*suite*)



Reconnaissance d'activités

- Par historique de mouvement (*suite*)
  - Pour les pixels ayant un $P_{\varsigma,\tau}(x,y,t)$ non nul, on peut déterminer si le changement est long terme ou moyen terme avec:

$$|I(x, y, t) - I(x, y, t-1)| > T_M$$

  - Si la différence est plus grande que T_M, le changement est moyen terme pour le pixel, i.e. sa valeur change.

Reconnaissance d'activités

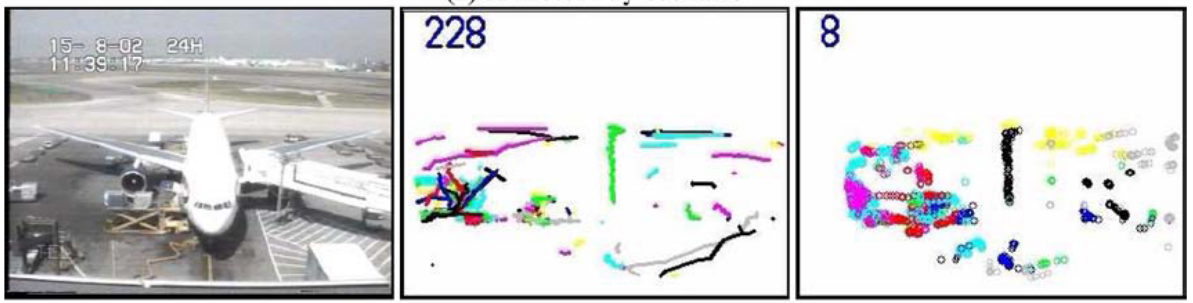
- Par historique de mouvement (*suite*)
  - Pour un blob, un événement est représenté par un vecteur de 7 caractéristiques (si $\overline{PCH} > T_b$):

$$v = [\bar{x}, \bar{y}, w, h, R_m, M_p x, M_p y]$$

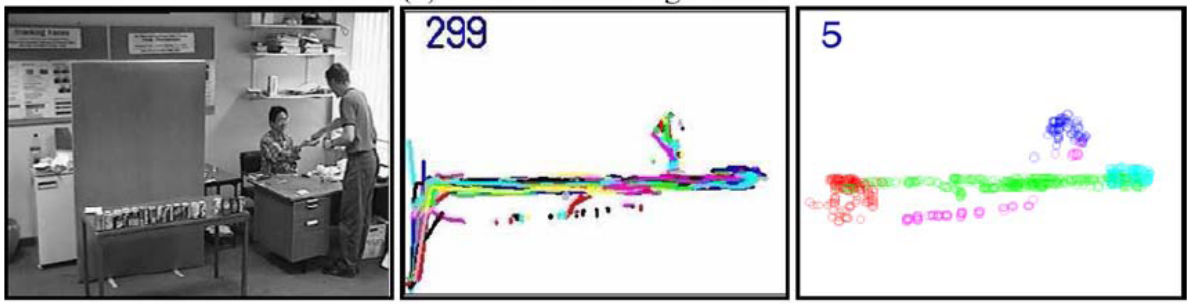
  - Où $(\bar{x}, \bar{y})$ est le centroïde, (w,h) la largeur et la hauteur de la forme, R_m est le pourcentage de pixels plus grand que T_M pour la forme, $(M_p x, M_p y)$ sont la direction du mouvement.
  - Ensuite, groupement et classification.

Reconnaissance d'activités

(a) A motorway scenario



(b) An aircraft docking scenario



(c) A shopping scenario

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - Basée sur des points distinctifs dans le domaine spatio-temporel.
  - Si on observe une action ou activité, et qu'on empile les images pour former un « cube » spatio-temporel, une action spécifique génèrera une collection de points spatio-temporels distinctifs.
  - Les points distinctifs sont calculés selon les mêmes principes en 3D qu'en 2D.

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - Exemples de points distinctifs 3D : HOG3D, LSS3D, SIFT3D, SURF3D, etc.
  - Au lieu de décrire une région *n x n*, ils décrivent des régions *n x n x n*.
  - Il existe aussi des descripteurs spatio-temporels qui ne sont pas 3D, exemple : MoSIFT.
  - Dans tous les cas, c'est toujours le même principe d'utilisation. Étudions le cas de MoSIFT.

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - MoSIFT (motion SIFT)
    - C'est un vecteur de 256 valeurs, les 128 premières sont les valeurs de SIFT sur les gradients dans une image autour du point, et les 128 dernières sont l'histogramme de l'orientation du flux optique dans l'image autour du point.
    - On extrait d'abord les points SIFT dans une trame décrivant une action. On calcule aussi le flux optique entre cette trame et une précédente.

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - MoSIFT (motion SIFT)
    - On décrit uniquement les points SIFT situés dans un endroit où le flux optique est non nul.

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - Apprentissage des actions
    - Pour toutes les actions d'intérêt, on extrait des points MoSIFT.
    - On fera ensuite des statistiques sur les vecteurs MoSIFT propres à une action donnée.
    - On calculera un histogramme des vecteurs MoSIFT pour chaque action.
    - Puisqu'il y a une très grande diversité dans les vecteurs MoSIFT, mais l'histogramme résultant épars, il faudra d'abord quantifier.

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - Apprentissage des actions
    - La quantification des vecteurs MoSIFT est faite par groupement. Il faut donc déterminer le nombre de groupements, et trouver les centres des groupements. Ensuite, chaque vecteur MoSIFT sera assigné à un groupe par sa distance par rapport aux centres.
    - Les centres des groupes sont habituellement obtenus avec l'algorithme *K-Means*.
    - On obtient des résultats proches en tirant les centres au hasard.

Reconnaissance d'activités

- Approche sac de mots visuels (bag-of-words)
  - Apprentissage des actions
    - Les centres constituent les mots visuels.
    - Pour chaque action on calcule la fréquence de chaque mot visuel, et le résultat est compilé dans un histogramme.
    - Pour améliorer les résultats, on aura plusieurs histogrammes pour la même action, et on apprendra à distinguer les histogrammes des actions à l'aide d'une méthode de classification comme *SVM (support vector machine)*.

Détection du transport d'objets

- Méthode basée sur la symétrie et sur des calculs de périodicité.
  - Périodicité: suppose des déplacements sans changements brusques d'orientation pour au moins 60 trames.
  - Symétrie: les objets doivent être suffisamment gros pour causer des dissymétries.

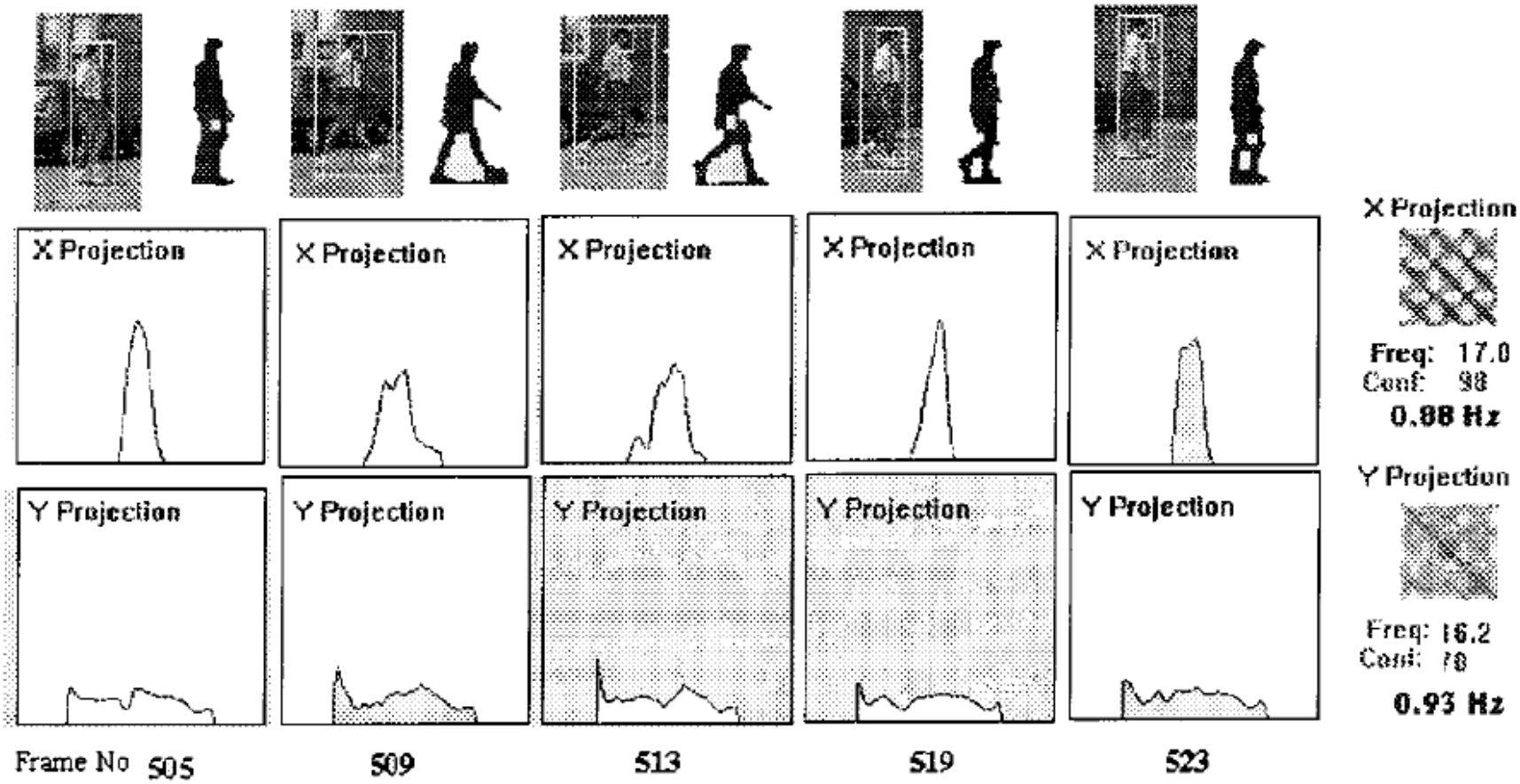
Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)
  - Calcul de la périodicité
    - On calcule les histogrammes de projection en X et en Y. Ils sont normalisés et alignés selon les axes de la personne (voir chapitre 6, pages 13-15).
    - Une matrice de similarité est ensuite créée. Soit p^{t1} et p^{t2}, deux histogrammes de projection aux temps t_1 et t_2. La similarité entre t_1 et t_2 est donnée par

$$S(t_1, t_2) = \min_{|dx|<q} \sum_{L \le i \le R} \left| p_{i+dx}^{t_1} - p_i^{t_2} \right|$$

où q est une fenêtre de recherche pour tenir compte des erreurs de suivi.

Détection du transport d'objets



Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)
  - Calcul de la périodicité
    - À partir de la matrice de similarité, une méthode d'auto-corrélation est appliquée pour détecter les périodes. La méthode utilisée est 1D et est appliquée ligne par ligne. L'auto-corrélation $R_r(\tau)$ pour un délai τ et une ligne r est calculée par:

$$R_r(\tau) = \sum_{t=0}^{N-|\tau|-1} S(r,t)S(r,t+\tau)$$

Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)
  - Calcul de la périodicité
    - la période T_r est un pic dans $R_r(\tau)$:

$$T_r = \text{argmax}\{R_r(\tau) : \tau \in [0, (N-1)F]\}$$

    - F est l'intervalle entre deux trames (0.033s pour vidéo NTSC). La fréquence ν_r est $1/T_r$. Parmi les pics ν_r pour chaque rangée, la fréquence fondamentale est celle la plus commune:

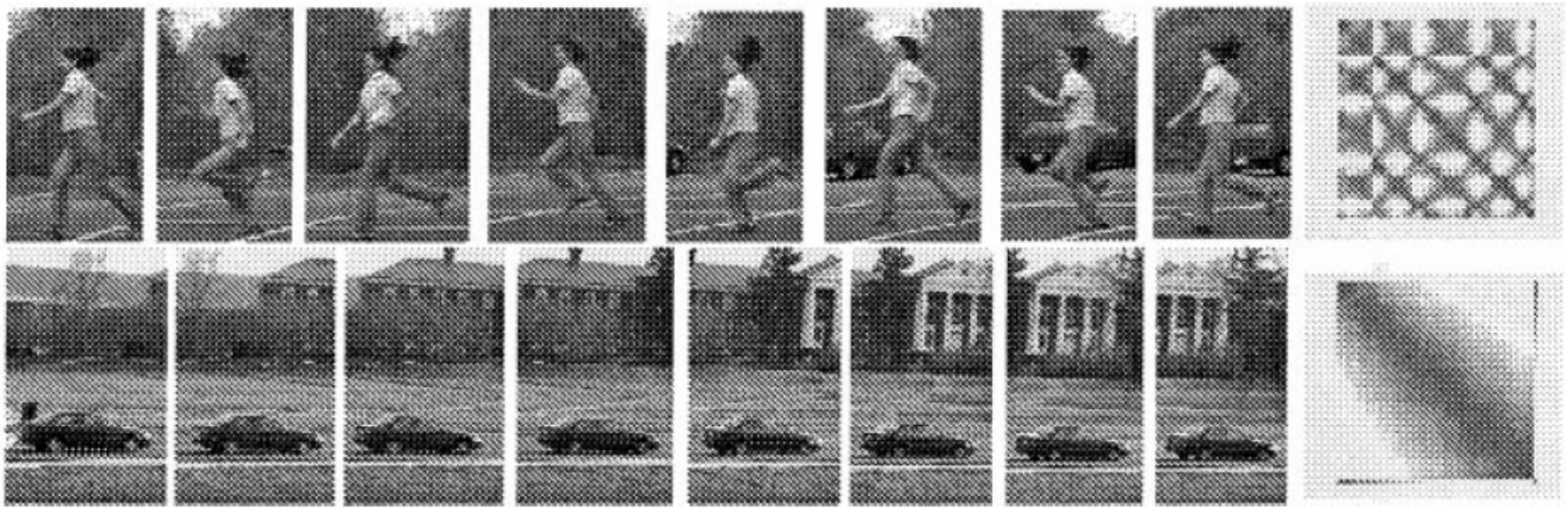
$$\nu = \text{argmax}\{C(f) : f \in [0, (N-1)F]\}$$

avec

$$C(f) = \frac{\sum_{r=0}^{N-1} \delta(\nu_r - f)}{N}$$

Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)



Détection du transport d'objets

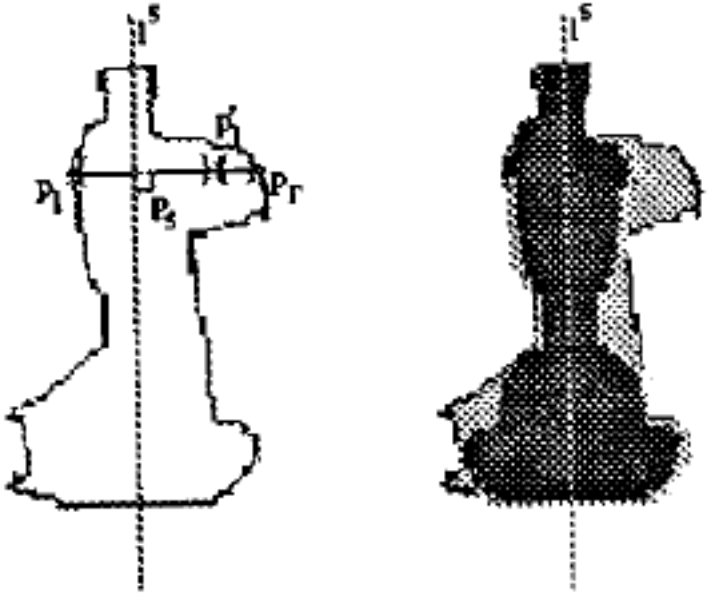
- Méthode basée sur la symétrie (*suite*)
  - Calcul de la symétrie
    - Soit q_i^l et q_i^r les longueurs entre les frontières et l'axe majeur pour une rangée i. Un pixel x est classé:

$$x = \begin{cases} \text{nonsymmetric} & \text{if } q_s^x > \min\{q_i^l, q_i^r\} + \epsilon \\ \text{symmetric} & \text{otherwise,} \end{cases}$$

où q_s^x est la distance entre l'axe majeur et un pixel et ϵ est une constante.

Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)

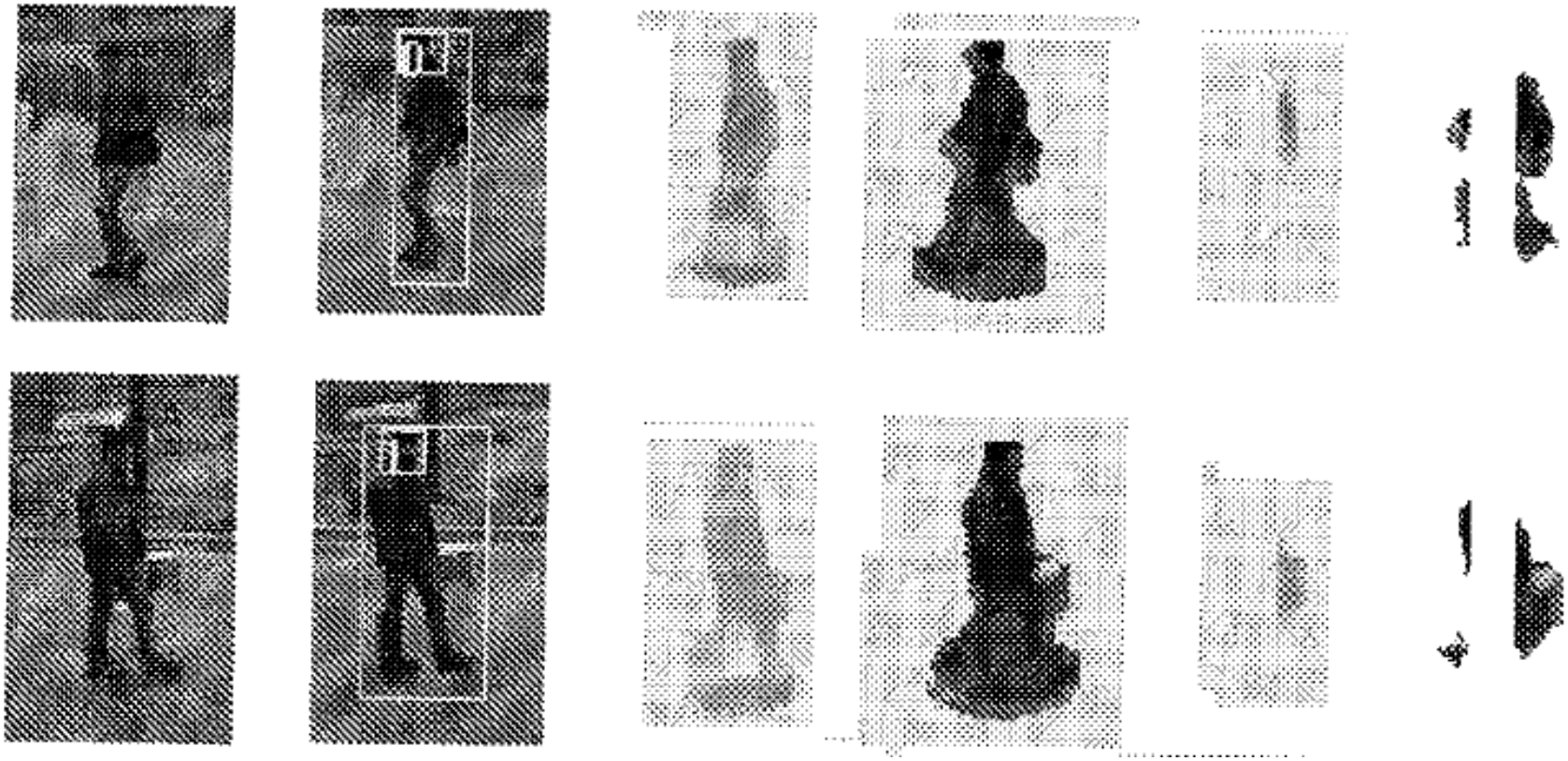


Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)
  - Pour détecter le transport d'objets, la première étape est le calcul de la symétrie. Les régions qui ne sont pas symétriques sont intégrées à un patron temporel de textures (Chapitre 6, page 16).
  - On peut alors trouver précisément les régions non-symétriques.

Détection du transport d'objets

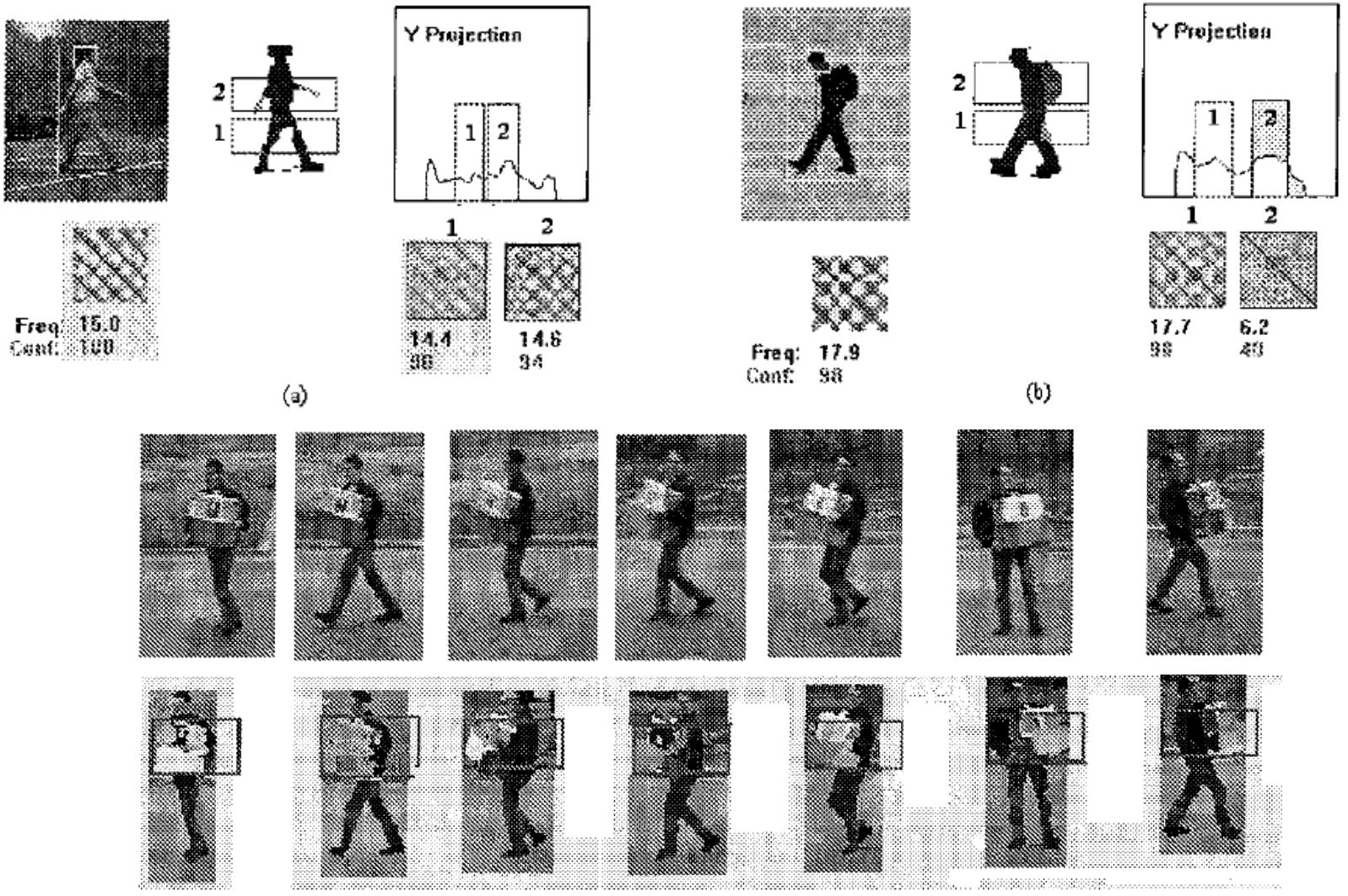
- Méthode basée sur la symétrie (*suite*)



Détection du transport d'objets

- Méthode basée sur la symétrie (*suite*)
  - Ensuite, on vérifie si les zones non-symétriques sont animées d'un mouvement périodique. Si ce n'est pas le cas, alors la zone non-symétrique contient un objet transporté.

Détection du transport d'objets



Bibliographie

- I. Haritaoglu et al., Backpack: Detection of People Carrying Objects Using Sihouettes, Computer Vision and Image Understanding 81, 2001, pp. 385–397
- L.M. Fuentes, S.A. Velastin, People Tracking in Surveillance Applications, Image and Vision Computing, 24, 2006, pp. 1165-1171
- Hongeng et al., Video-based event recognition: activity representation and probabilistic recognition methods, Computer Vision and Image Understanding ,96, 2004, pp. 129–162
- C. Lu, N.J. Ferrier, Repetitive Motion Analysis: Segmentation and Event Classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 2, 2004, pp. 258-263
- D. Makris, T. Ellis, Learning Semantic Scene Models From Observing Activity in Visual Surveillance, IEEE Transactions on Systems, Man, And Cybernetics – Part B: cybernetics, Vol. 35, No. 3, 2005, pp. 397-408
- T.Xiang, S. Gong, Beyond Tracking: Modelling Activity and Understanding Behaviour, International Journal of Computer Vision 67(1), 2006, pp. 21–51
- M.Y. Chen, A. Hauptmann, MoSIFT: Recognizing Human Actions in Surveillance Videos, Technical report, CMU-CS-09-161, Carnegie Mellon University, 2009