Chapitre 1

Détection de mouvement

Sommaire

1.1 Introduction

La détection de mouvement dans les séquences d'images (vidéo) a comme objectif de trouver les points (pixels) qui changent de position au cours du temps dans la séquence, c'est à dire consiste d'associer à chaque pixel une étiquette binaire pour modéliser le changement (valeur 1) ou le non changement de position (valeur 0). Le résultat final de la détection de mouvement est donc une carte binaire qui contient des valeurs 1 et 0, ces dernières peuvent être utilisées pour identifier la présence ou l'absence de mouvement dont le but est de distinguer les objets mobiles et statiques dans la scène (voir la Figure 1.1).

FIGURE 1.1 – Exemple de la détection de mouvement avec la séquence «Cube de Rubik»

La Figure 1.1 représente la séquence cube de Rubik (de taille 240×256 pixels) qui est constituée d'un cube placé au centre d'un plateau circulaire et qui tourne sur lui-même avec une

vitesse constante, donc les zones avec couleur blanche représentent le mouvement, ce dernier est dû au déplacement des objets dans la scène. Le processus de détection de mouvement n'est pas toujours possible à cause de plusieurs obstacles, nous pouvons citer :

- La variation de la luminosité et les conditions d'éclairage qui ont des influences sur l'apparence des objets composant la scène.
- Bruit dans la séquence d'images.
- Variabilité de la forme et de la structure des objets de la scène.
- Les zones homogènes où nous ne pouvons pas détecter le mouvement bien que ce dernier existe.

Par conséquent, les méthodes de détection de mouvement sont généralement basées sur la contrainte suivante : l'intensité et l'éclairement de la scène doivent être constants (ou avec de petites variations) au cours du temps. Une autre contrainte que nous pouvons ajouter est que le capteur soit immobile, cette contrainte est utile puisque le processus de détection de mouvement dans les séquences d'images est la première étape dans plusieurs domaines d'application, par exemple :

- La segmentation basée sur l'étude de mouvement
- L'interprétation des séquences d'image pour pouvoir donner des décisions à partir des images.
- Le suivi d'objet.
- L'analyse de comportement.
- La compression et la prédiction vidéo
- La vidéo-surveillance
- La robotique pour éviter les collisions

1.2 Détection basée sur la différence entre images

La méthode intuitive pour la détection de mouvement dans une séquence d'images est de détecter la différence pixel par pixel c'est-à-dire mesuré le changement temporel des pixels dans la séquence, ce qui correspond au calcul du gradient temporel en chaque pixel. Aussi, si dans la séquence on un mouvement nul, il n'y aura pas de variation entre les images correspondantes ; si par contre dans un mouvement, un changement évident se produira entre les images correspondantes.

Parmi les méthodes de détection de mouvement basées sur la différence entre les images on y trouve la méthode de différence ou de soustraction entre deux images successives. Dans une séquence d'images au niveaux de gris, et pour chaque pixel de coordonnées (x, y) dans l'image ou frame I_{t-1} qui correspond a l'image a l'instant $t-1$, on calcule la différence absolue avec ses coordonnées correspondantes dans l'image suivante (acquise a l'instant t) notée I_t comme suit :

$$\zeta(x, y) = |I_t(x, y) - I_{t-1}(x, y)| \tag{1.1}$$

Pour une image couleur RGB par exemple, on peut calculer cette différence absolue par, la distance Euclidienne comme suit :

$$\zeta(x, y) = \sqrt{(I_t^R(x, y) - I_{t-1}^R(x, y))^2 + (I_t^G(x, y) - I_{t-1}^G(x, y))^2 + (I_t^B(x, y) - I_{t-1}^B(x, y))^2} \tag{1.2}$$

Où $I_t^R(x, y)$, $I_t^G(x, y)$ et $I_t^B(x, y)$ représentent l'intensité du pixel de coordonné (x, y) dans les champs R, G et B respectivement du système de couleurs RGB.

En présence de mouvement entre les deux images alors $\zeta \neq 0$ et dans le cas contraire c'est-à-dire $\zeta = 0$ on un mouvement nul. Puisque les images présentent en générale un bruit lié par exemple au capteur dans les systèmes de surveillance ou d'interprétation (dans la phase d'acquisition), une opération de seuillage est indispensable. Le seuillage permet d'une part d'avoir une information pertinente et d'autre part de produit comme résultat une image binaire qui indiquent les régions de mouvement (seules correspondantes aux variations de l'intensité lumineuse dans les deux images), formellement on peut écrire :

$$\Psi(\zeta(x,y)) = \begin{cases} 1 & \text{si } \zeta(x,y) > \tau, \\ 0 & \text{sinon} \end{cases} \tag{1.3}$$

où τ est le seuil, qui décide de la sensibilité de la détection de mouvement.

La méthode de détection de mouvement par la différence entre deux images successives est caractérisée par sa simplicité, sa facilité de mettre en œuvre, sa flexibilité, son adaptation au changement de la scène, ainsi cette méthode demande peu de ressources. Comme inconvénient on peut citer la sensibilité au bruit et ne traite pas le problème des zones homogènes.

Exemple :La figure 1.2 montre un exemple de la méthode de différence entre deux images successive :

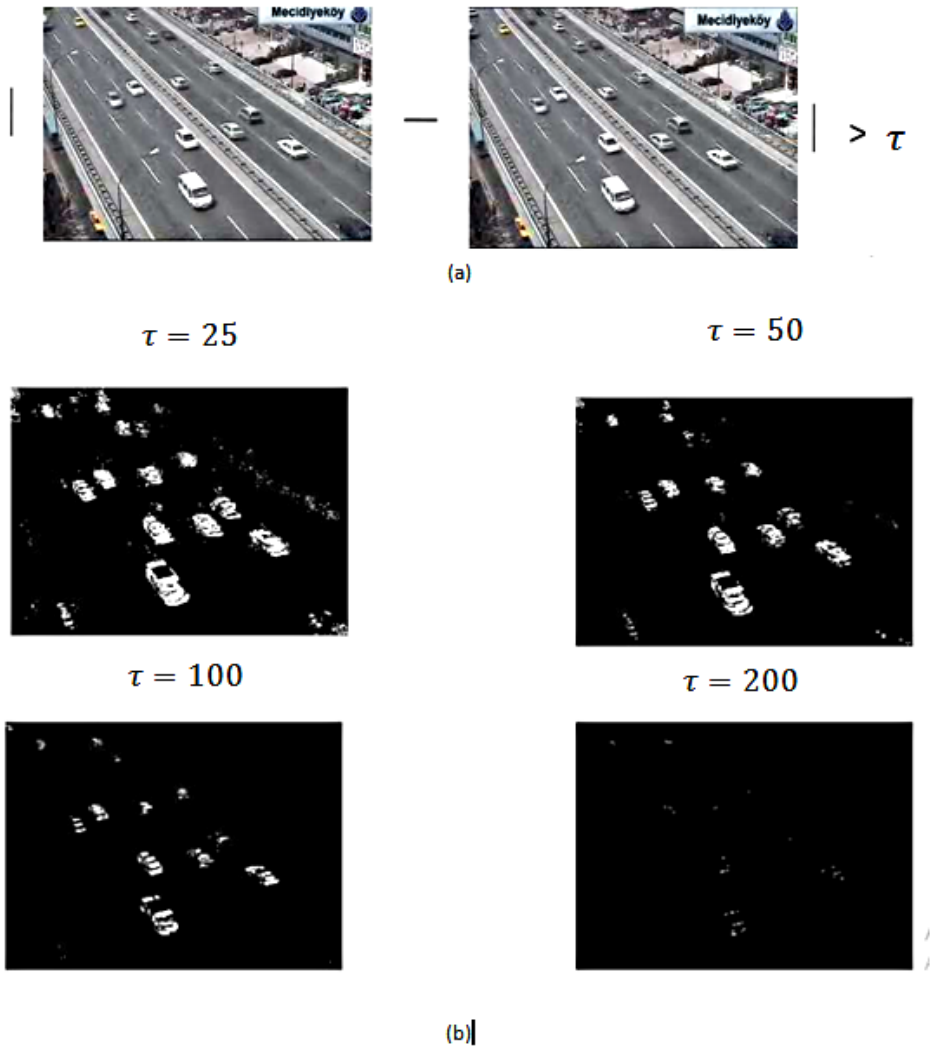

FIGURE 1.2 – Exemple :(a) principe de la détection, (b) résultats obtenues

Pour l'amélioration de la méthode précédente qui utilise une seule opération de différence entre les deux trames, une alternative de cette méthode est proposée, la méthode proposée utilise deux opérations de différences entre trois trames pour traiter le problème de faible mouvement et aussi le problème de bruits. La méthode de différence entres trois trames successives noté $I_{(t-1)}$, I_t et $I_{(t+1)}$ consiste à mesuré la différence entre les trois trames selon les équations 1.4 et 1.5 ensuite une opération de seuillage des résultats obtenus selon les équations 1.6 et 1.7 et enfin combiner les résultats obtenus précédemment selon l'équation 1.8 comme suit :

$$\zeta_1(x,y) = |I_t(x,y) - I_{t-1}(x,y)| \tag{1.4}$$

$$\zeta_2(x,y) = |I_{t+1}(x,y) - I_t(x,y)| \tag{1.5}$$

$$\Psi(\zeta_1(x,y)) = \begin{cases} 1 & \text{si } \zeta(x,y) > \tau_1, \\ 0 & \text{sinon} \end{cases} \tag{1.6}$$

$$\Psi(\zeta_2(x,y)) = \begin{cases} 1 & \text{si } \zeta(x,y) > \tau_2, \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

$$\zeta(x,y) = \min(\Psi(\zeta_1(x,y)), \Psi(\zeta_2(x,y))) \tag{1.8}$$

Exemple :

Dans la Figure 1.3 illustre un exemple de la méthode de différence avec 3 trames exécutives avec la séquence cube de Rubik.

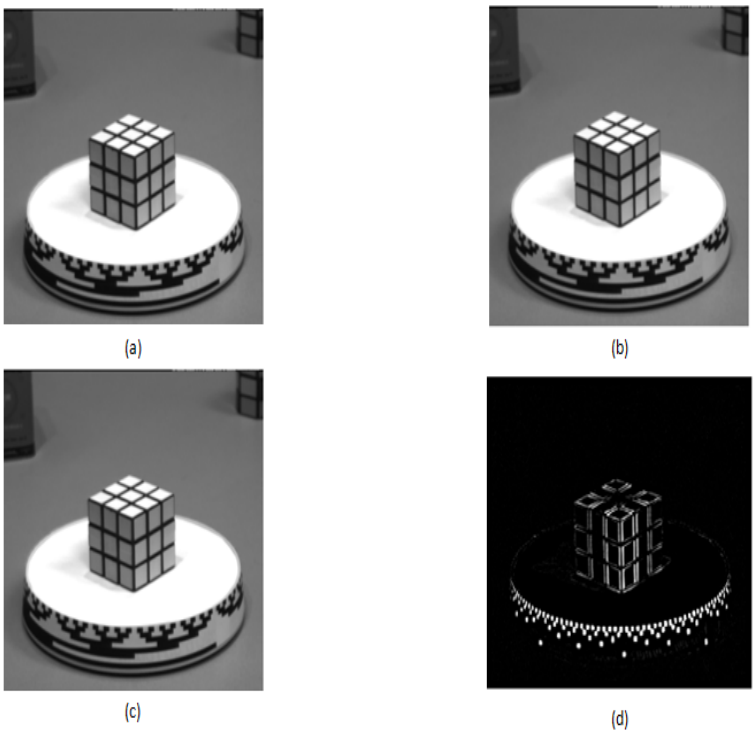

FIGURE 1.3 – Exemple :(a) 1ère image, (b) 2ème image,(c) 3ème image,(d) résultat

Pour surmonter le problème des zones homogènes qui imposent que les images soient bien texturées, une autre méthode qui est basée sur la différence entre les trames appelée la détection de mouvement avec image de référence est proposée. La méthode proposée requière l'utilisation d'une image de référence notée I_R, cette image doit représente l'arrière plan sans aucun objets mobile c'est-à-dire seuls les objets immobiles sont présents dans cette image, aussi l'image de référence doit prendre en compte le changement de l'arrière plan dans chaque image de la séquence, donc l'image de référence doit être régénérée à chaque Δt(ce qui correspond au temps d'acquisition pour chaque image de la séquence). Donc, la qualité des résultats obtenus est liée avec la qualité de l'image de référence. On distingue deux techniques pour la construction de l'image de référence :

- Soit on sélectionne nous même l'image de référence, on prend par exemple :la première image dans la séquence, ou l'image qui contient peu/aucun mouvement.
- Soit on construite l'image de référence comme suit :

$$I_R = \sum_{i=t-N}^{t-1} w_i I_t \tag{1.9}$$

avec : $w_i = \frac{1}{N}$ où : I_R est l'image de référence
I_t : L'image a l'instant t
N : Le nombre de frame dans ma séquence
w_i Représente le facteur de pondération.

La détection de mouvement avec image de référence consiste à mesurer la différence entre l'image de référence et l'image courante selon l'équation 1.10 et enfin une opération de seuillage des résultats obtenus précédemment pour réduire le bruit :

$$\zeta(x, y) = |I_t(x, y) - I_R(x, y)| \tag{1.10}$$

Cette méthode est moins rapide par rapport aux méthodes vu précédemment puisqu'elle nécessite la reconstruction de l'image de référence à chaque Δt, aussi elle est sensible aux variations considérables de luminosité au cours du temps surtout pour les séquences qui sont composée de nombre important des images, c'est-à-dire que l'image de référence doit avoir les même variations de luminance qui existe dans la séquence d'images.

Exemple :

Un exemple de la méthode de détection avec image de référence est illustré sur la Figure 1.4 :

FIGURE 1.4 – Exemple :(a) image de la séquence, (b) image de référence,(c) résultat(Seuil absolu= 25)

Une autre technique a été proposée par A. Bello [1], qui utilise à la fois les avantages des deux méthodes décrites précédemment (la soustraction des trois images successives et la méthode de détection avec image de référence) est définie comme suit :

— Calculer le détecteur à court terme entre les images $I_{(t-1)}$, I_t et $I_{(t+1)}$ qui permet d'avoir la différence entre les trois trames comme suit :

$$\zeta^1(x,y) = \min(\zeta_1(x,y), \zeta_2(x,y)) \tag{1.11}$$

— Calculer le détecteur à long terme entre l'image de référence I_R et l'image I_t selon l'équation 1.12

$$\zeta^2(x,y) = |I_t(x,y) - I_R(x,y)| \tag{1.12}$$

— Calculer la différence d'image robuste entre les résultats des étapes précédentes comme suit :

$$\zeta^1(x,y) = \max(\Psi(\zeta^1(x,y)), \Psi(\zeta^2(x,y))) \tag{1.13}$$

Exemple :

Un exemple de la méthode de détection de A. Bello est illustré sur la Figure 1.5 avec la même séquence (routière) que l'exemple précédant :

FIGURE 1.5 – Exemple :(a) image de la séquence, (b) résultat obtenu

Les méthodes de détection de mouvement basées sur la différence d'image soufrent du problème de bruit dans la séquence d'images malgré tous les améliorations reconnues, cela

1. Aldo. Bellon. Détection et suivi de véhicules en mouvement dans une séquence d'images. PhD thesis, Université Blaise Pascal, Clermond-Ferrand, Octobre 1996.

provoque des coupures irrégulier dans les zones de mouvement détectées et donc des trous dans des objets mobiles. Ce problème peut être surmonté par une opération de lissage (en utilisant un filtre spatial Gaussien par exemple) avant le seuillage. Aussi, Ce mécanisme de détection de mouvement est limité généralement dans les régions qui ne présentent aucune variation de texture (ou on a un faible gradient d'image).

1.3 Test de vraisemblance

Parmi les méthodes de détection de mouvement on trouve les méthodes qui exploitent le contenu stochastique de la séquence d'image, ces méthodes sont connues sous le nom des méthodes probabiliste. Dans cette section nous présentons la détection de mouvement au sens du maximum de la vraisemblance [?] ou la fonction(test) de vraisemblance c'est-à-dire détection par test des hypothèses statistiques.

Le test de vraisemblance est connu comme un test des hypothèses qui est utilisé pour comparer deux modèles en établissant une hypothèse dite hypothèse nulle, et en utilisant certaines données pour rejeter l'hypothèse dans le cas où la vraisemblance sous l'hypothèse alternative est significativement supérieure à la vraisemblance sous l'hypothèse nulle et réciproquement. Dans notre domaine de la détection de mouvement, on applique le test de vraisemblance comme suit[?] :

Considérons I_t et $I_{(t+dt)}$ les deux trames que l'on cherche à détecter le mouvement entre eux, ces deux trames sont découpée en blocs ou zones notées A_1, A_2 respectivement centré sur le pixel p de coordonné (x, y) et soit $g_i^1(i = 1, \ldots, m)$, $g_j^2(j = 1, \ldots, n)$ des valeurs observées dans les deux zones A_1 et A_2 qui correspondent à des valeurs de niveaux de gris corrompues par un bruit Gaussien de moyenne 0 et de variance σ^2 noté $N(0, \sigma)$ (voir la Figure 1.6).

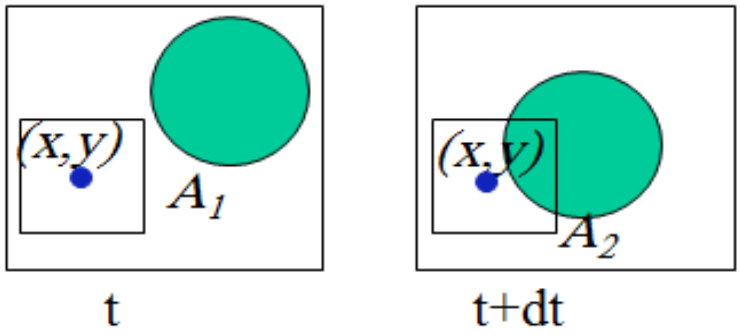

FIGURE 1.6 – Détection par max de vraisemblance

La distribution de probabilité conjointe pour l'observation de ces mesures dans le cas de n valeurs d'échantillon est donnée par la fonction de vraisemblance suivante :

$$L = f(g, \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(g_i - \mu)^2\right) \tag{1.14}$$

où μ représente la moyenne de la distribution des valeurs d'échantillon.

Les hypothèses à tester sont notées comme suit :

- $H0$: A_1 et A_2 proviennent de la même distribution $N(\mu_0, \sigma_0) \Longrightarrow$ pas de mouvement(zone stationnaire)
- $H1$: A_1 et A_2 proviennent de différentes distributions $N(\mu_1, \sigma)$ et $N(\mu_2, \sigma) \Longrightarrow$ le mouvement existe(zone mobile)

Si $H1$ vraie, la fonction de vraisemblance sera (Loi de Gauss) :

$$L(H1) = \left(\frac{1}{2\pi\sigma^2}\right)^{\left(\frac{m+n}{2}\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{m}(g_i{}^1 - \mu_1)^2 + \sum_{i=1}^{n}(g_i{}^2 - \mu_2)^2\right)\right] \tag{1.15}$$

Et les dérivées partielles par rapport à σ^2, μ_1, μ_2 seront comme suit :

$$\begin{cases} \hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m} g_i^1, \\ \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} g_i^2, \\ \hat{\sigma}^2 = \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \left(\frac{1}{m+n}\right)\left[\sum_{i=1}^{m}\left(g_i^1 - \hat{\mu}_1\right)^2 + \sum_{i=1}^{n}\left(g_i^1 - \hat{\mu}_2\right)^2\right]. \end{cases} \tag{1.16}$$

Si $H0$ vraie, la fonction de vraisemblance sera $L(H0)$:

$$\begin{cases} \hat{\mu}_0 = \frac{1}{m+n}\left[\sum_{i=1}^{m} g_i^1 + \sum_{i=1}^{n} g_i^2\right], \\ \hat{\sigma}_0^2 = \left(\frac{1}{m+n}\right)\left[\sum_{i=1}^{m}\left(g_i^1 - \hat{\mu}_0\right)^2 + \sum_{i=1}^{n}\left(g_i^1 - \hat{\mu}_0\right)^2\right]. \end{cases} \tag{1.17}$$

Alors, le rapport de vraisemblance maximum peut être écrit sous la forme suivant :

$$\lambda = \frac{L(H1)}{L(H0)} \tag{1.18}$$

Donc on choisi $H1$ si $\lambda > 1$ et $H0$ dans le cas contraire, formellement le critère de décision s'écrit comme suit :

$$\begin{cases} H1 & \text{si } \lambda > 1 \\ H0 & \text{si } \lambda \leq 1 \end{cases} \tag{1.19}$$

Pratiquement le rapport de vraisemblance maximum peut être écrire sous la forme suivant [?] :

$$\lambda = \left[1 + \frac{t^2}{m+n-2}\right]^{\left(\frac{m+n}{2}\right)} \tag{1.20}$$

et nous pouvons tester t^2 au lieu de λ puisque cette dernière est une fonction monotone de t^2, donc on sélectionne un seuil t_α, en fonction d'un niveau de confiance α

Avec :

$$t = \frac{\left(\frac{mn}{m+n}\right)^{\frac{1}{2}}\left(\hat{\mu}_1 - \hat{\mu}_2\right)}{\left[\frac{\sum_{i=1}^{m}\left(g_i^1 - \hat{\mu}_1\right)^2}{m+n-2} + \frac{\sum_{j=1}^{n}\left(g_i^2 - \hat{\mu}_2\right)^2}{m+n-2}\right]^{\frac{1}{2}}} \tag{1.21}$$

Cette méthode de Hsu [?] et qui utilise le test de vraisemblance ne permet pas de prendre en compte une variation d'illumination, pour cela des améliorations ont été proposées, alors :

— Pour un changement d'illumination global, une transformation linéaire est utilisé pour que la moyenne (μ) et la variance (σ) de l'image normalisée soient respectivement égales à la moyenne et à la variance de l'image d'origine comme suit :

$$\hat{I}_t(g) = aI_t(g) + b \tag{1.22}$$

tel que :

$$\hat{\mu}_t = \mu_{t+dt}$$

et

$$\hat{\sigma}_t^2 = \sigma_{t+dt}^2$$

alors :

$$\hat{I}_t(g) = \frac{\sigma_{t+dt}}{\sigma_t}\left(I_t(g) - \mu_t\right) + \mu_{t+dt} \tag{1.23}$$

— Pour des changements d'illumination locaux, on fait la comparaison des gradients d'intensité, comme suit

$$\frac{1}{N}\sum_{g\in A} \parallel \nabla I_t(g) - \nabla I_{t+dt}(g) \parallel_{\leq}^{\geq} \theta \tag{1.24}$$

Où θ représente la seuil.

Exemple [2] :

Si $m = n = 12$, pour un niveau de confiance $\alpha = 5\%$, puis $t_{0.05} = 1.717$ et $t_{0.05}^2 = 2.948$;pour niveau de confiance $\alpha = 1\%$, on obtient $t_{0.01} = 2.508$ et $t_{0.01}^2 = 6.29$. Lorsque la valeur de t^2 est inférieur à t_α^2 on choisi l'hypothèse $H0$. Sinon, $H1$ sera accepté.

1.4 Détection basée sur un modèle

Les méthodes de détection de mouvement basées sur la différence entre les images souffrent de plusieurs problèmes, on trouve trop de pixels en mouvement si le seuil est bas plus la complexité des algorithmes, aussi le bruit de capteur (acquisition+numérisation) dans les applications en temps réel. Alors la solution est d'utiliser les méthodes de relaxation par modèle (ou champs) de Markov (MRF : Markov Random Fields). A cet effet, nous présentons dans cette section la méthode proposée par Bouthemy et Lalande [**?**], cet algorithme vise à améliorer la différence d'image par l'utilisation d'un processus markovien.

Pour l'approche Markovienne, la solution du problème de la détection de mouvement est la réalisation la plus probable d'un certain phénomène aléatoire, l'image est considérée comme un ensemble de site noté S où chaque site note s_i représente un pixel de l'image et à chaque site est associé un descripteur qui peut être son niveau de gris ou son étiquette, donc un champ aléatoire X sur S à valeur dans V est défini par :

$X : \Omega \rightarrow E^S$

Où :

Ω : Univers probabiliste

2. Hsu Y. Z, NAGEL H.- H., et REKERS G.," New Likelihood Test Methods for Change Detection in Image Sequences" . COMPUTER VISION, GRAPHICS, AND IMAGE PROCESSING 26, 73 − 106(1984).

E^S : ensemble des valeurs de sites.

Dans notre contexte de la détection de mouvement on a : $\begin{cases} E = \{0,1\} & \text{(fixe, mobile)} \\ S = Z^3 & \text{espace-temps discret} \end{cases}$

L'ensemble des sites S est muni d'un système de voisinage noté v qui détermine les relations de dépendances entre les variables aléatoires X_s et dans notre cas de la détection il modélise les relations spatio-temporelles entre les divers sites de la séquence d'images, formellement on a :

$v : S \rightarrow P(S)$

$$\forall (s,r) \in S \begin{cases} s \notin v(s) \\ s \in v(r) \Longrightarrow r \in v(s) \end{cases} \tag{1.25}$$

A partir d'un système de voisinage v, un système de cliques est défini. Une clique est soit un singleton de S, soit un ensemble de sites de tous voisins les uns des autres. Une clique est dite d'ordre K, si elle contient K éléments. on peut distingué trois types de cliques : clique spatiale, clique temporelle et clique spatio-temporelle (voir la Figure 1.7).

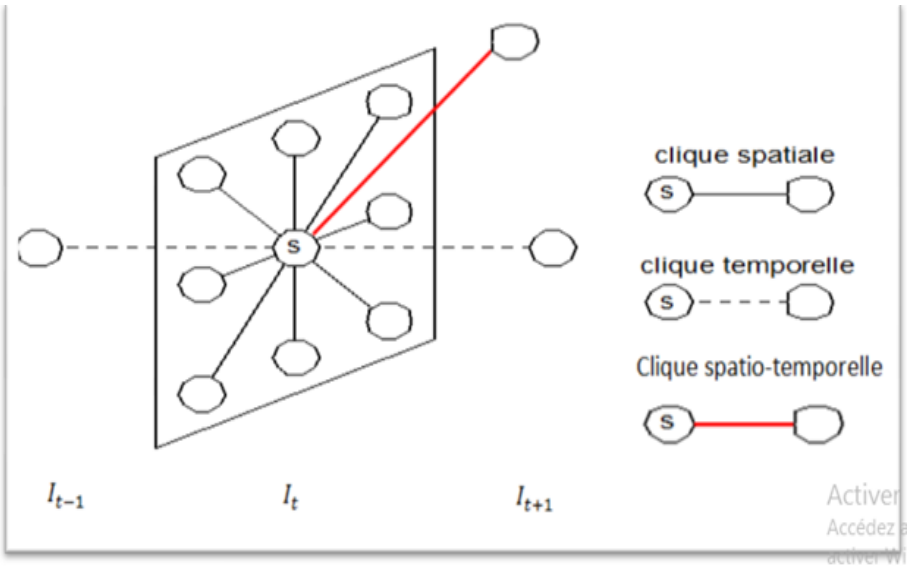

FIGURE 1.7 – système de cliques

L'image est modélisée de façon probabiliste comme une réalisation d'un champ aléatoire, alors selon le modèle markovien « X est un champ de Markov si et seulement si la probabilité conditionnelle locale en un site n'est fonction que de la configuration du voisinage du site considéré ». En d'autre terme, on recherche la réalisation la plus probable d'un champ de Markov dit « caché » X, à valeur binaire sur Z^3 : les étiquètes, à partir d'un champ connu dit « observation » O, qui correspond à la séquence des différences d'images (équation 1.26).

$$O_t(x,y) = \mid I_t(x,y) - I_{t-1}(x,y) \mid \tag{1.26}$$

L'estimateur Bayesien du Maximum A Posteriori est utilisé par les auteurs pour estimer les étiquettes appropriées du champ E menu du champ d'observation O qui consiste a la maximisation de la probabilité conditionnelle comme suit :

$$\max \frac{P[O = o \mid E = e]P[E = e]}{P[O = o]} \tag{1.27}$$

La maximisation du numérateur de l'équation 1.27 est équivalente à la minimisation d'une fonction d'énergie dérivée du théorème de Hammersley-Clifford, qui stipule que les champs aléatoires de Markov suivent la distribution, ce qui nous a permet de décrire un modèle de champ de Markov dans une topologie donnée en spécifiant les potentiels attachés à chaque clique comme suit :

$$P[O = o \mid E = e] = \frac{e^{-U(o,e)}}{Z} \tag{1.28}$$

où Z est un facteur de normalisation. La fonction énergétique U est donnée par la somme de deux termes comme suit :

$$U(e, o) = U_m(e) + U_a(o, e) \tag{1.29}$$

Avec U_m représente l'énergie qui assure l'homogénéité spatio-temporelle et U_a désigne l'énergie d'adéquation qui assure une bonne cohérence de la solution par rapport aux données observées :

$$U_a(o, e) = \frac{1}{2\sigma^2}[o - \psi(e)]^2 \tag{1.30}$$

avec :

$$\psi(e) = \begin{cases} \text{o} & \text{si} \mid I_t(x, y) - I_{t-1}(x, y) \mid < \tau \\ \alpha & \text{sinon} \end{cases}$$

et

$$U_m = \sum_{c \in C} V_c(e_s, e_r)$$

De plus, V_c est donné selon le modèle de Potts [?] qui permet de prendre en compte différentes relations entre entre différentes valeurs des descripteurs : étiquettes.

$$V_c(e_s, e_r) = V_s(e_s, e_r) + V_p(e_s^t, e_s^{t-1}) + V_p(e_s^t, e_s^{t+1}) \tag{1.31}$$

avec : $V_s(e_s, e_r) = \begin{cases} -\beta_s & \text{si } e_s = e_r \\ +\beta_s & \text{sinon} \end{cases}$, $V_p(e_s^t, e_s^{t-1}) = \begin{cases} -\beta_s & \text{si } e_s^t = e_r^{t-1} \\ +\beta_s & \text{sinon} \end{cases}$ et

$V_p(e_s^t, e_s^{t+1} = \begin{cases} -\beta_s & \text{si } e_s^t = e_r^{t+1} \\ +\beta_s & \text{sinon} \end{cases}$

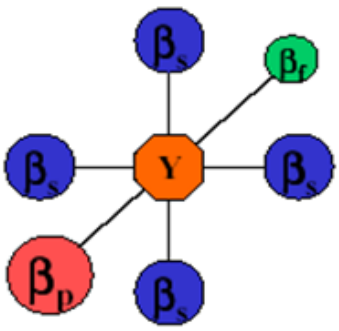

FIGURE 1.8 – Modèle de Potts

Lorsque β est positif, les configurations les plus probables correspondent à des sites voisins de même niveau de gris ou descripteur.

Après avoir défini l'énergie U pour le modèle markovien, les auteurs ont considéré le problème de la minimisation de cette énergie.

Exemple : la Figure 1.9 représente un exemple de la méthode décrite ci-dessus avec deux séquences d'images (Sofa et Canoe) et les paramètres suivants : $\beta_s = 20, \beta_p = 10, \beta_f = 30$ et $\alpha = 10$.

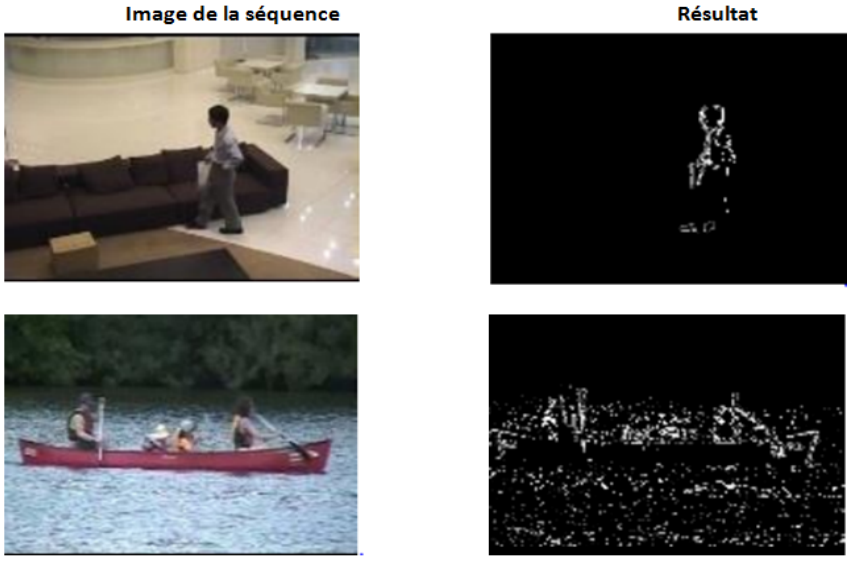

FIGURE 1.9 – Exemple avec dela méthode de Bouthemy et Lalande