IDIAP RESEARCH REPORT

ANALYSE NON SUPERVISéE D'ACTIVITéS EN VIDéO SURVEILLANCE POUR L'ANALYSE DE SCèNE ET LA DéTECTION D'éVéNEMENTS ANORMAUX

Remi Emonet Jean-Marc Odobez

Idiap-RR-20-2013

MAY 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Analyse non supervisée d'activités en vidéo surveillance pour l'analyse de scène et la détection d'événements anormaux

Rémi Emonet et Jean-Marc Odobez

21 mai 2013

1 Introduction

L'installation de caméras de vidéo surveillance a connu une croissance importante ces dernières années. L'exploitation de ces caméras réponds à de nombreux besoins et à différents objectifs : un objectif de sûreté, lorsqu'il s'agit d'assurer l'intégrité physique des personnes dans un environnement, par exemple lors de la montée ou descente de passagers dans un train ou un métro, ou pour détecter des incidents pouvant conduire à des accidents en environnement autoroutier ou urbain ; un objectif de sécurité et de protection des équipement, par détection d'intrusions, de bagages abandonnés, d'agressions et en général de tout comportement anti-social ou de vandalisme ; un objectif d'efficacité, par identification de tendances de flux afin de détecter des congestions (en milieu routier par exemple) et les prévenir par des moyens appropriés – information et recommandations aux usagers, modifications d'itinéraires, etc.

Néanmoins, dans une grande majorité des cas, les caméras sont de simples boites d'enregistrement, dont les données ne sont exploitées qu'à posteriori lorsqu'un crime a été commis. Ceci résulte de la volonté générale ou de raisons légales (afin de protéger la vie privée), mais également de facteurs techniques et économiques : les algorithmes d'analyse automatique ne sont pas suffisamment fiables, ou les coûts d'infrastructure liés à l'emploi de tels algorithmes ou au recourt à des opérateurs de surveillance sont trop élevés compte tenus du nombre important de caméra à analyser. Au delà des leur performance, un facteur important limitant l'utilisation d'algorithme concerne leur configuration lors du déploiement en environnement réel : elle est généralement technique, difficile, et effectuée par des non-spécialistes du domaine, c'est-à-dire sans connaissance vis-à-vis des problèmes liés à la vision par ordinateur.

Dans ce contexte, un effort important est actuellement dirigé vers la recherche d'algorithmes capable de découvrir à partir d'observations d'une heure à plusieurs jours les activités typiques d'une scène, lorsqu'elles commencent et se terminent, les relations entre elles, les moments où elles sont le plus susceptibles de se produire, etc. Une telle information peut être utile en elle même, afin de mieux comprendre le contenu de la scène et sa dynamique, ou en prétraitement avant une analyse de plus haut niveau. Par exemple, l'analyse permettrait de dériver les véritables activités du point de vue capteur, fournir un contexte pour d'autres tâches (par exemple le suivi ou l'interprétation de données) ou de définir des indicateurs de situations anormales qui pourrait être exploités pour sélectionner automatiquement les vues à présenter à un opérateur dans une salle de contrôle surveillant plusieurs centaines de caméras.

Dans ce chapitre nous présentons une classe récente d'approches allant dans ce but. Basées sur des modèles dits « à thème » (« topic or theme models » en anglais) par référence à leurs développements initiaux dans le domaine de l'analyse sémantique de textes, ces approches non supervisées ou faiblement supervisées permettent de découvrir les activités principales d'une scène, les cycles éventuels, les anomalies, par analyse des co-occurrences de mots visuels. Ces derniers sont définis par quantification de caractéristiques simples de la vidéo – comme la position dans l'image (et indirectement dans la scène), le mouvement apparent, des indices de taille ou de forme – qui s'extraient de manière immédiates et évitent de ce fait le recours au suivi des objets dans la scène, une tâche actuellement difficile en pratique pour des environnements encombrés.

Le contenu du chapitre sera organisé comme suit. Dans une première partie, nous étudierons tout d'abord le principaux concepts des modèles à thèmes au travers de l'un de ses modèles les plus simple : Probabilistic Latent Semantic Analysis, PLSA. Nous montrerons ensuite dans cette même partie comment celui-ci peut-être appliqué à la recherche d'activités dans des vidéos grâce à la définition d'un vocabulaire (choix des mots et ce qu'ils représentent) et de documents appropriés. Dans une deuxième partie, nous présenterons un modèle plus récent que nous avons proposé, et qui permet de découvrir des thèmes temporels (appelés motifs par la suite), i.e. des thèmes qui ne capturent pas simplement la co-occurrence des mots à un instant donné, comme le fait PLSA, mais aussi l'ordre dans lesquels les mots se produisent au cours du temps. Les deux parties seront illustrées de résultats visualisant les activités découvertes. Dans une troisième partie, nous fournirons des exemples d'utilisation de ces modèles, par exemple pour la détection des anomalies ou la prédiction, ainsi que des pistes pour leur évaluation. Le chapitre se terminera sur une discussion des travaux actuels et des futurs travaux à envisager dans ce domaine.

2 Un exemple de modèle à thème : PLSA

2.1 Introduction

Récemment, le design de modèles probabilistiques bayesiens appelés modèles à thèmes est devenu une direction de recherche pertinente pour découvrir des patterns récurrents dans des données issues de toutes sortes de capteurs. Ces modèles proviennent à l'origine du domaine qui traite de l'analyse automatique de textes. Ils considèrent un texte comme un sac-de-mots (SDM ; bag-of-words en anglais) obtenu en comptant le nombre d'occurrence de chaque mot dans le document, éliminant ainsi toute information sur leur ordre d'apparition. Malgré cette simplification, comme les mots capturent une information sémantique substantielle, les SDM sont une représentation qui a été utilisé avec succès pour de nombreuses tâches de l'analyse de texte, comme la classification en différent genres ou la récupération de texte (text retrieval). Les modèles à thèmes, comme PLSA [Hof01] ou le modèle LDA (Latent Dirichlet Allocation [BNJ03]), sont construits à partir des SDM, et ont été introduits pour découvrir les thèmes dominants dans des collections de données par analyse de la co-occurrence des mots, une notion analogue à la corrélation mais qui s'applique à des données discrètes.

Grâce à leur pouvoir d'analyse, leur facilité d'implémentation, leur versatilité, et leur nature non-supervisée, les modèles à thème ont été appliqués à un grand de nombre de problèmes et de modalités comme outil de fouille de données. En particulier, ils ont été utilisés sous différentes formes pour découvrir des activités humaines à partir de vidéos de sport [NWL08] ou de surveillance [WMG09], de données d'accéléromètres [HFS08], ou des mesures GPS de téléphones mobiles [FGP08]. Néanmoins, la spécification d'un vocabulaire et de documents appropriés pour la capture des activités d'intérêts, la modélisation effective des informations spatiales et temporelles, l'interprétation des résultats et la détection des événements non usuels restent des défis aussi bien généraux que pour un domaine d'application donné.

Dans la section qui suit, nous présentons plus en détail le modèle PLSA, et expliquons ensuite comment il peut-être appliqué à des vidéos pour découvrir des activités.

2.2 Modèle PLSA

Le modèle PLSA [Hof01] a été introduit comme une version probabiliste de l'analyse sémantique latente (LSA) pour capturer les information co-occurrentes récurrentes dans un ensemble de données discrètes. Bien que considéré comme un modèle non entièrement génératif, la simplicité de son modèle d'optimisation

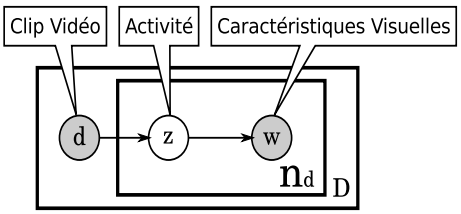

FIGURE 1 – Le modèle génératif PLSA. Les nœux ombrés représentent des variables observées, alors que les autres dénotent des variables latentes.

en fait une alternative intéressante vis-à-vis des modèles entièrement génératifs comme LDA [BNJ03].

Processus génératif – Le modèle graphique est donné en figure 1. Pour un ensemble de documents d, le processus de génération des observations (les couples (w,d) de mots apparaissant dans les documents et formant les SDM) est le suivant :

- tirer aléatoirement le document d dans lequel l'observation sera générée selon une probabilité $p(d)$;
- tirer le motif $z \sim p(z|d)$, où $p(z|d)$ représente la probabilité qu'occupe le thème z dans le document d, c'est à dire indirectement, la probabilité qu'un mot w du document appartienne au thème z ;
- tirer le mot $w \sim p(w|z)$, où $p(w|z)$ est la probabilité que le mot w apparaisse dans le thème z.

Comme le montre ce processus, PLSA associe à chaque observation (w, d) une variable latente $z \in \mathcal{Z} = \{z_1, \ldots, z_{N_z}\}$ définissant le thème du mot observé. La probabilité jointe du modèle résultant est alors donnée par :

$$p(w, z, d) = p(d)p(z|d)p(w|z) \tag{1}$$

D'un point de vue probabiliste, cela introduit une hypothèse d'indépendance conditionnelle entre les variables observées, à savoir que l'apparition d'un mot est indépendante du document étant donné le thème auquel il appartient. Avec ce modèle, la vraisemblance des observations est donc :

$$p(w, d) = p(d)p(w|d) = p(d) \sum_{z=z_1}^{z_{N_z}} p(z|d)p(w|z). \tag{2}$$

Comme l'indique cette expression, le modèle décompose la distribution $p(w|d)$ des mots dans un document en une combinaison linéaire convexe des thèmes $p(w|z)$, où les poids $p(z|d)$ sont fournis par la distribution des thèmes dans le document. On a donc un modèle de mélange typique, à l'image des mélanges de gaussiennes pour des données réelles.

Inférence – L'estimation des paramètres Θ (les différentes tables de proba-

bilité, i.e. $\Theta = \{p(d), p(z|d), p(w|z)\}$) se fait typiquement suivant le principe du maximum de vraisemblance. Plus précisément, étant donné un ensemble de données d'entraînement $\mathcal{D}$, la log-vraisemblance de Θ s'exprime par :

$$\mathcal{L}(\Theta|\mathcal{D}) = \sum_{d \in \mathcal{D}} \sum_{w} \mathrm{n}(d, w) \log \left(p(w, d) \right) \tag{3}$$

où $p(w, d)$ est donnée par l'équation 2. En pratique, compte tenu de la présence de la somme dans le logarithme, l'optimisation est conduite à l'aide d'un algorithme EM (Expectation-Maximization) itératif standard dans lequel les probabilités des variables cachées sont estimées puis utilisées dans une étape de maximisation des paramètres [Hof01]. Cette procédure conduit à l'estimation des thèmes $p(w|z)$ et des distributions des thèmes $p(z|d)$ dans les documents d'apprentissage.

Nouveaux documents – Dans ce cas, nous sommes seulement intéressés par l'estimation des poids $p(z|d)$ des thèmes dans le nouveau document d. Ceux-ci s'obtiennent en utilisant le même algorithme EM que ci-dessus, mais sans mise à jour des thèmes $p(w|z)$, et qui conduit simplement à la maximisation de la log-vraisemblance normalisée $\mathcal{L}^{norm}$ dans chaque document d :

$$\mathcal{L}_d^{norm}(p(z|d)) = \frac{1}{\mathrm{n}_d} \sum_{w} \mathrm{n}(d, w) \log \left(\sum_{z} p(z|d) p(w|z) \right) \tag{4}$$

où $\mathrm{n}_d = \sum_w \mathrm{n}(d, w)$.

2.3 PLSA appliqué aux vidéos

Le modèle PLSA s'applique à n'importe quel type de données. En analyse vidéo, nous souhaitons que les thèmes découverts caractérisent les activités fréquentes de la scène. En pratique la sémantique des thèmes dépendra essentiellement de la définition du vocabulaire et de la manière dont les documents sont construits. Dans cette section, nous présentons un exemple simple de construction de vocabulaire et de documents, et illustrons les résultats obtenus.

Vocabulaire – Celui-ci doit caractériser le contenu de la scène, et est obtenu par quantification de caractéristiques simples de la vidéo. L'exemple typique de la littérature s'appuie sur deux caractéristiques : la position, et le mouvement.
Position. Dans les vidéos de surveillance, la plupart des activités sont caractéristiques de l'endroit où elles se produisent. De ce fait, il est intéressant d'en tenir compte pour la construction du vocabulaire. La position est souvent quantifiée en cellules (ou blocs) de 4×4 à 10×10 pixels.
Mouvement. C'est une information essentielle pour différencier les activités.

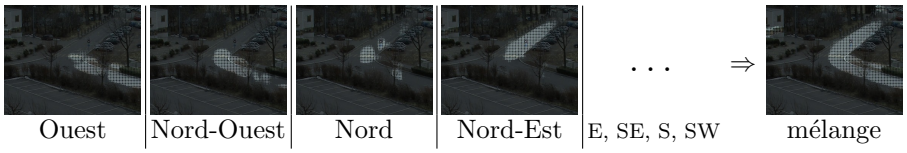

FIGURE 2 – Re-projection d'un thème (fenêtre de 20 secondes) sur chacune de ses 8 orientations puis dans une version condensée mélangeant les 8 orientations. Les 4 orientations omises (E, SE, S, SW) ne contiennent dans cet exemple aucune activité.

L'estimation peut être faite de manière robuste (e.g., à l'aide de l'algorithme de Lucas-Kanade multi-résolutions) vis-à-vis des variations de contenu (texture), et avec un coût calcul réduit. De manière intéressante, elle est relativement indépendante des conditions d'illumination. Pour être utilisé comme mot, le mouvement doit être quantifié. Classiquement, la direction est jugée plus importante, et les mouvement d'amplitude suffisante sont classés par quantification de leur direction en 4 ou 8 étiquettes.

Vocabulaire. Nous définissons le vocabulaire comme le produit cartésien des espaces position et mouvement. Ainsi, pour une image de 280×360 pixels, et en utilisant des blocs de taille 4×4, on aboutit à un total de $70 \times 90 \times 8 = 50400$ mots potentiels. En pratique, lors de l'apprentissage, cet ensemble peut-être réduit en éliminant tout les mots qui n'apparaissent jamais ou qui représentent moins de 0.5% des observations. Au final, on obtient en général de 10000 à 20000 mots.

Documents – Ceux-ci sont simplement construits en divisant la vidéo en clips vidéo de courtes durées. On obtient alors le SDM de chaque document d en comptant le nombre de fois $\mathrm{n}(d, w)$ qu'il contient chaque mot w.

Exemple de thèmes découverts – La méthode PLSA est illustrée en considérant une vidéo de 1h45min de la scène visible sur la figure 3. Dans ce cas, l'activité d'un véhicule peut-être décrit comme un ensemble de mouvements (position et orientation) qui co-occurrent dans le clip vidéo. Chaque activité correspond donc à un thème représenté par la distribution $p(w|z)$ des caractéristiques visuelles qui co-occurrent souvent.

Pour identifier les positions actives d'un thème donné, on peut, pour chaque position c, marginaliser cette distribution par rapport à l'ensemble des mots V_c rattachés à cette position (et ayant différentes orientations de mouvement). On obtient alors une carte d'activité : $p(\text{activité} \in c|z) = \sum_{w \in V_c} p(w|z)$. La figure 2 illustre cette marginalisation à l'échelle de l'image complète.

Les images de la figure 3 montrent quelques thèmes retrouvés lorsque la

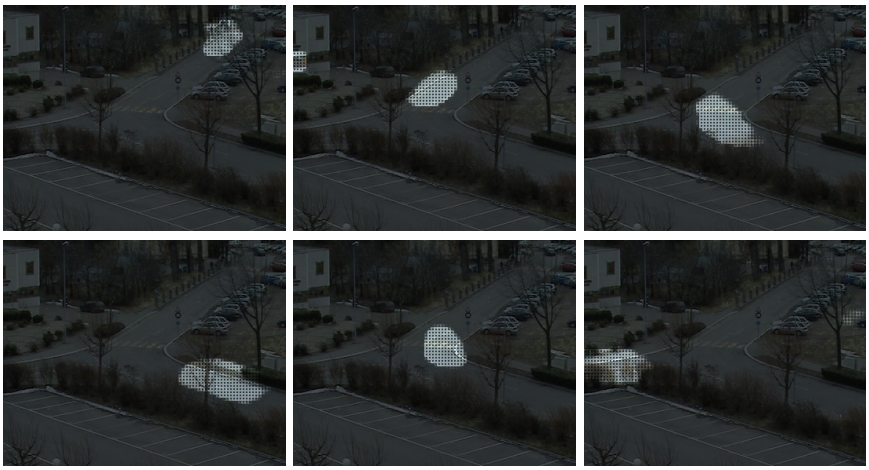

FIGURE 3 – Exemples de thèmes obtenus (6 sur 75) en utilisant comme document des clips vidéo de durée 1 seconde.

durée des clips et de 1 seconde et que l'on cherche à retrouver $N_z = 75$ thèmes. Ceux-ci correspondent effectivement aux activités élémentaires qui peuvent se produire en une seconde, et permet de reconstruire par combinaison linéaire l'activité totale observée pendant une telle durée.
Afin de capturer des activités plus sémantiques, on peut allonger la durée des clips. Les résultats lorsque celle-ci est de 10 secondes (et $N_z = 20$) sont présentés sur la figure 4. Bien que les activités capturées sur cet exemple aient du sens, on peut noter que l'information temporelle est perdue, et que la détermination du moment où une telle activité se produit ne sera pas aisée.

Influence du vocabulaire – Les exemples précédents on mis en avant l'influence de la durée sur les thèmes retrouvés. On pourrait par ailleurs constater, en visualisant l'ensemble des résultats l'absence de thèmes représentant les stops des voitures au carrefour. En effet, une telle information n'est capturée par aucun mot. Dans [VO09], ceci est pris en compte en appliquant une étape de soustraction de fond. Ceci permet de créer des mots avec l'étiquette statique : ce sont les points de l'avant plan dont le mouvement estimé est nul. On retrouve alors après application de PLSA des thèmes spécifiques liés à cette caractéristique. Notons que dans ce même article, l'utilisation de mots liés à la taille des blobs obtenus par soustraction de fond permet également de clairement distinguer (sur une autre scène) les activités de piétons de celles des voitures, notamment sur les passages piétons.

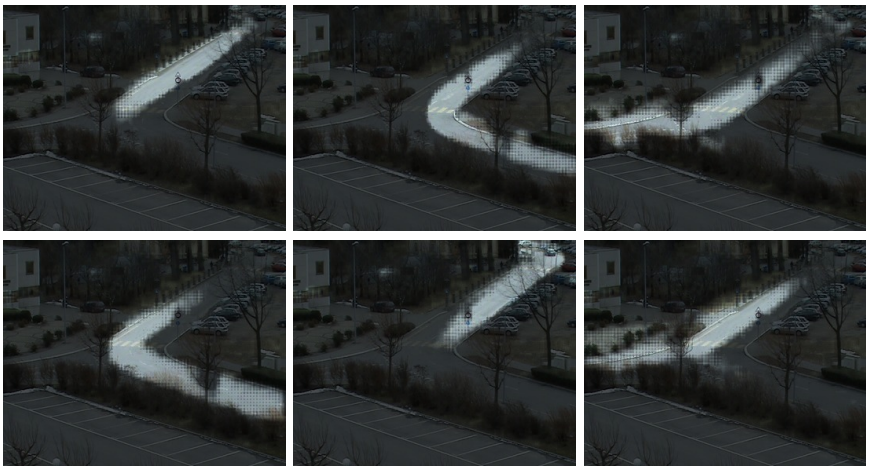

FIGURE 4 – Quelques thèmes obtenus (6 sur 20) en utilisant comme document des clips vidéo de durée 10 secondes. PLSA permet bien de retrouver les principales activités de la scène. Il est à noter que l'information temporelle est perdue.

3 PLSM et Modèles Temporels

Comme expliqué dans la section précédente, les modèles à thèmes simples comme PLSA permettent de capturer les activités récurrentes d'une scène mais perdent toute information temporelle à l'intérieur d'un document. Le modèle PLSM (≪ Probabilistic Sequential Motifs ≫ en anglais) permet de pallier ce problème et de capturer l'information temporelle dans un thème que l'on nomme alors ≪ motif ≫. D'autres modèle essaient de modéliser le temps mais, contrairement à PLSM, sont incapables d'à la fois séparer les activités récurrentes et de trouver quand celles ci apparaissent. Cette section est dédiée à la présentation du modèle PLSM.

3.1 Modèle PLSM

Le modèle PLSM prend en entrée un ensemble de documents temporels défini par une matrice de comptes $n(w, t_a, d)$ formé par une accumulation d'observations, chacune étant un triplet (w, t_a, d). Comme illustré par la figure 5a, PLSM modélise un document temporel comme une combinaison d'éléments latents :

– d'une part, des motifs (2 dans l'exemple) : chaque motif z est une distri-

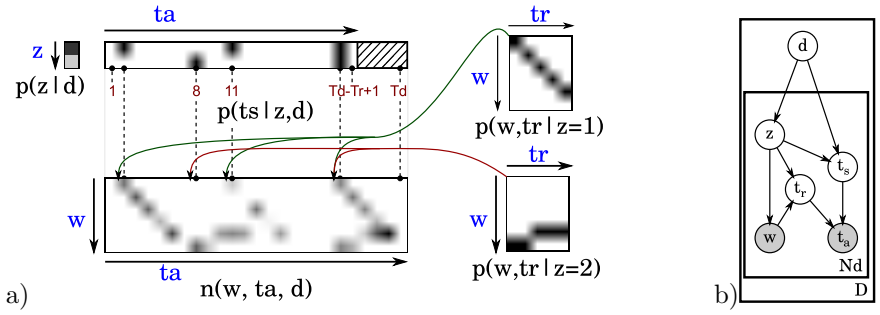

FIGURE 5 – Processus Génératif – a) génération d'un document temporel d ; b) modèle graphique (les observations sont grisées).

bution $p(w, t_r|z)$ définie sur le produit cartésien du vocabulaire et d'un axe temporel,

– d'autre part, les instants dans le document temporel où les motifs apparaissent : chaque motif z démarre selon une table $p(t_s|z, d)$.

Processus génératif – Le modèle graphique de PLSM est donné dans la figure 5b. Pour un ensemble de documents temporels, le processus de génération de chaque observation (w,t_a,d) est le suivant :

– tirer le document d dans lequel l'observation sera générée,

– tirer le motif : $z \sim p(z|d)$, où $p(z|d)$ est la probabilité qu'une observation du document d provienne d'un motif z,

– tirer un instant d'occurrence : $t_s \sim p(t_s|z, d)$, où $p(t_s|z, d)$ est la probabilité qu'un motif z commence à l'instant t_s dans le document d.

– tirer un mot et un temps relatif $(w, t_r) \sim p(w, t_r|z)$ [1], où $p(w, t_r|z)$ est la probabilité qu'un mot w apparaisse à un temps t_r dans un motif z,

– affecter $t_a = t_s + t_r$ (t_a est totalement défini sachant t_s et t_r).

La distribution jointe sur l'ensemble des variables du modèle peut être dérivée du processus génératif. Étant donnée la relation déterministe $t_a = t_s + t_r$, uniquement deux de ces trois variables apparaissent généralement dans les équations. La distribution jointe pour le modèle PLSM est la suivante :

$$p(w, t_a, d, z, t_s, t_r) = p(d)\, p(z|d)\, p(t_s|z, d)\, p(w, t_a - t_s|z) \tag{5}$$

Inférence – Le but final du modèle PLSM est d'extraire, à partir de documents temporels, les motifs et quand ils apparaissent. Ces éléments sont les paramètres du modèle, notés Θ, et composés de $p(z|d)$, $p(t_s|z, d)$ et $p(w, t_r|z)$.

1. dans la figure 5b le tirage de (w,t_r) est décomposé en un tirage de t_r suivi du tirage de w sachant t_r

L'estimateur du maximum de vraisemblance peut être obtenu en maximisant la log-vraisemblance des données observées (notées $\mathcal{D}$). Après marginalisation sur les variable cachées $Y = \{t_s, z\}$, la log-vraisemblance s'écrit :

$$\mathcal{L}(\Theta|\mathcal{D}) = \sum_{d=1}^{D} \sum_{w=1}^{N_w} \sum_{t_a=1}^{T_d} n(w, t_a, d) \log \sum_{z=1}^{N_z} \sum_{t_s=1}^{T_{ds}} p(w, t_a, d, z, t_s, t_r) \qquad (6)$$

Un algorithme d'espérance-maximisation (EM) peut être dérivé de l'expression de la log-vraisemblance et permet d'obtenir les paramètres du modèle. Les détails de cette procédure EM sont disponibles dans [VEO10]. Pour améliorer la qualité des résultats obtenus, il est possible d'intégrer une contrainte de parcimonie (*sparsity* en anglais) directement au sein de l'algorithme EM.

3.2 Motifs Extraits par PLSM

Dans cette section, nous illustrons les résultats obtenus par PLSM appliqués sur des vidéos.

Création de documents temporels – Pour des raisons de coût de calcul, PLSM n'est pas directement appliqué sur les documents présentés dans la section 2. Pour simplifier les observations, les documents de bas niveau sont d'abord traité avec PLSA en utilisant un nombre de thèmes relativement important, par exemple 75. La réponse de chaque thème de PLSA à un instant donné est utilisé comme observation pour PLSM. De cette façon, la taille du vocabulaire est réduite de plusieurs milliers à seulement 75. De la même façon, une réduction de la résolution temporelle est réalisée : une fenêtre d'une seconde est utilisée pour réduire la fréquence à 1 observation par seconde. Au final, pour une vidéos de une heure, la table $n(w, t_a, \cdot)$ serait de taille 75×3600.

Représentation des motifs – Un motif est une table donnant une probabilité pour chaque mot du vocabulaire à chaque instant relatif. Les mots du vocabulaire correspondent à des thèmes PLSA et donc à des régions d'activité dans l'image. À chaque instant relatif, il est donc possible de re-projeter les différents mots du vocabulaire dans l'image. Une animation montrant successivement les instants relatifs permet de voir le motif dans le temps. En utilisant un dégradé de gris, il est possible de condenser cette animation en une seule image pour la représentation sur papier. La figure 6 illustre ces différentes représentations.

Exemples de Motifs Extraits par PLSM – Les figures 7 et 8 montrent les motifs représentatifs obtenus sur deux scènes différentes. D'une manière générale, PLSM est à même d'extraire les activités principale dans les scènes qui lui sont données.

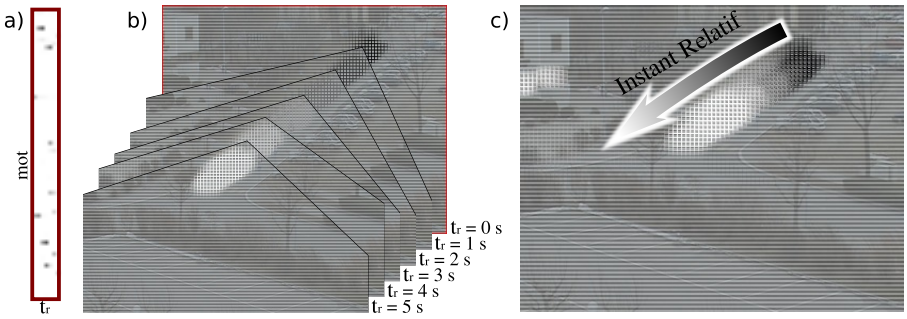

FIGURE 6 – Différentes représentations pour les motifs – a) sous forme de table ; b) en re-projetant chaque instant relatif t_r ; c) en utilisant un dégradé de gris pour représenter le temps.

FIGURE 7 – Jeu de donné d'un carrefour avec piétons et voitures – ligne supérieur : activités de véhicules ; seconde ligne : activités de piétons.

3.3 Applications : comptage, anormalité, etc.

PLSM apporte une compréhension rapide de la scène les motifs grâce aux motifs qu'il extrait. Au delà de cette compréhension, il est possible d'utiliser les motifs et leurs instants d'apparition de différentes façons. Nous montrons dans cette section comment le modèle peut être utilisé pour faire d'une part du comptage et d'autre part de la détection d'anormalité. Notons qu'une fois les motifs appris par PLSM, il est possible de fixer ces motifs et de trouver quand ils apparaissent dans une nouvelle vidéo.

Comptage – Les motifs extraits par PLSM correspondent à des activités récurrentes de la scène. Quand un motif z correspond à un événement donné, il est possible d'utiliser les instants d'apparition de ce motif pour construire

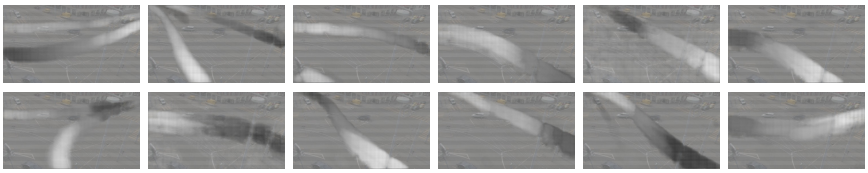

FIGURE 8 – Jeu de donnés d'un carrefour complexe avec voitures et tramways. Les différentes activités de voitures et de tramways sont extraites correctement.

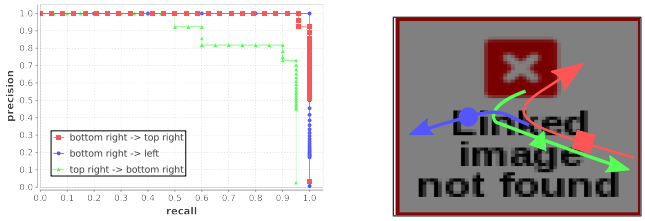

FIGURE 9 – Courbe de précision/rappel pour 3 types d'événements associés à 3 motifs sur 12 minutes de vidéo.

un détecteur pour l'événement correspondant. Il suffit d'appliquer un seuillage sur $p(t_s|z,d)$ pour créer un détecteur d'événement. Nous avons réalisé des détecteurs de cette façon et évalués sur des vidéos pour lesquelles nous avons annoté la vérité terrain. La figure 9 donne les courbes de précision/rappel pour 3 types d'événements. Les courbes sont obtenues en faisant varier la valeur du seuil appliqué à $p(t_s|z,d)$. Les résultats sont très bons et PLSM est à même de détecter les événements fréquents qu'il a appris. Des résultats équivalents ont été obtenus dans le cadre de la détections d'événements audio.

Détection d'anormalité – PLSM capture les activités récurrentes et donc les motifs représentent les choses normales (courantes). Si l'on a déjà appris des motifs pour une scène, il est possible d'étudier à quel point ces motifs sont capables d'expliquer une nouvelle vidéos de cette scène. Pour ce faire, PLSM est utilisé pour trouver les instants d'apparition des motifs (fixés) dans la nouvelle vidéo. D'autres mesures peuvent être envisagées (comme la vraisemblance) mais nous utilisons ici l'erreur de reconstruction pour quantifier à quel point la nouvelle vidéo est explicable à partir des motifs. L'anormalité basée sur l'erreur de reconstruction est définie ainsi :

$$abnormality(t_a,d) = \sum_w \left| \frac{n(w,t_a,d)}{n(d)} - p(w,t_a|d) \right| \tag{7}$$

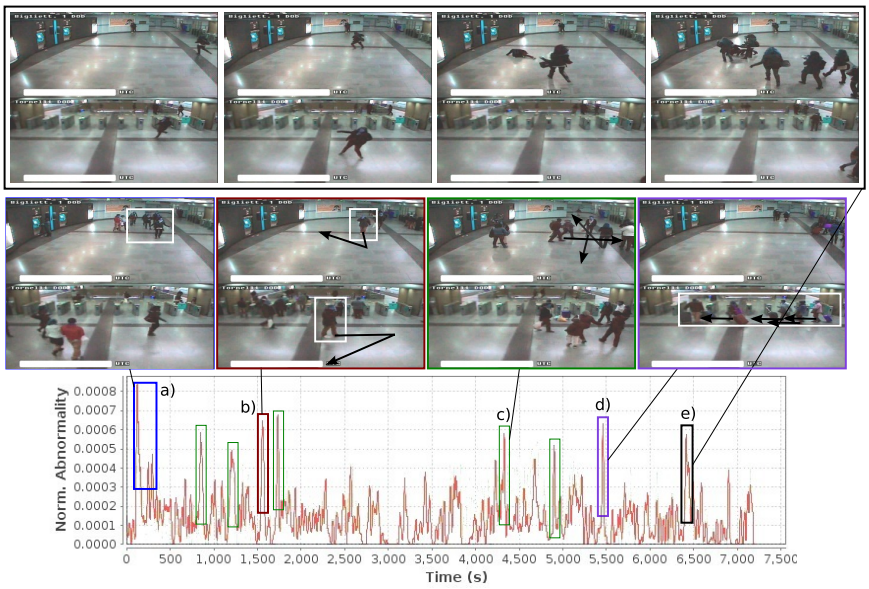

FIGURE 10 – Détection d'anormalité dans une station de métro

où : $p(w, t_a|d) = \sum_{t_s} \sum_z p(t_s, z|d)p(w, t_r = t_a - t_s|z)$

L'erreur de reconstruction est calculée à chaque instant t_a. Un seuillage permet de créer un détecteur d'anormalité. La figure 10 donne une courbe d'anormalité produite par PLSM appliqué sur une paire de caméra d'une station de métro. Les pics d'anormalité dépassant un seuil sont illustrés par des captures des instants correspondants. Une partie des anormalités détectées sont dues à des groupes bougeant in habituellement (Fig. 10a et Fig. 10d). Une grande partie sont causées par des trajectoires inhabituelles de personnes dues à un encombrement de la station (Fig. 10c ainsi que tous les autres rectangles en trait fin sur la courbe). Une importante anormalité est détectées (Fig. 10e) : une personne arrive en courant en suivant une trajectoire circulaire puis tombe et est rejointe par un groupe qui lui vient en aide.

Il est intéressant de noter que PLSA (sans aspect temporel) permet déjà de détecter de nombreuses anormalités quand cette détection ne requiert pas de raisonnement temporel. Cette détection d'anormalités en utilisant PLSA a été étudiée en détail dans [VO09] où différentes mesures d'anormalités sont comparées comme illustré dans la figure 11.

Sélection de capteur – Au delà de la pure détection d'anormalité, la mesure d'anormalité peut aussi être utilisée pour pré-sélectionner les capteurs (caméras

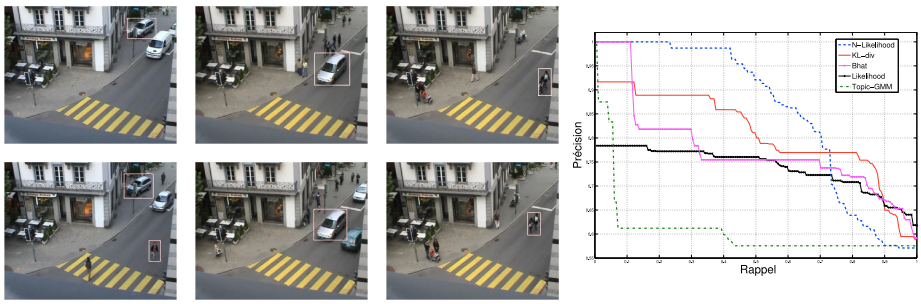

FIGURE 11 – Anormalité à partir de PLSA – Six exemples d'anormalité dans la scène considérée et courbes de précision/rappel obtenues à partir de différentes mesures d'anormalités (issues de PLSA).

et microphones) à soumettre pour interprétation à un opérateur. En effet, la plupart des caméras installées dans un réseau de métro ne peuvent être visualisées en continu faute de personnel ; une pré-sélection des caméras ayant un fort degré d'anormalité est alors très avantageuse pour améliorer la sécurité.

Prédiction et statistiques – Les modèles incorporant une information temporelle comme PLSM peuvent être utilisés pour réaliser une prédiction à court terme des activités observées : quand une activité est commencée, il est possible de supposer comment elle se finira. Une analyse statistique des apparitions des activités capturées par PLSM permet aussi de réaliser des prédiction à plus long terme. Par exemple l'utilisation d'une station de métro peut être étudiée et prédite, les modèles à thèmes servant ici à extraire des descripteurs de haut niveau (e.g., les occurrences d'une activité typique).

4 Conclusion

Ce chapitre a présenté les approches à base de « sac de mots » et de « modèles à thèmes ». D'une manière générale, ces approches sont particulièrement adaptées et efficaces pour réaliser une extraction non supervisé des activités principales contenues dans une scène.

À travers la sélection d'un vocabulaire (pour les mots, e.g., position et orientation du mouvement) et d'un modèle à thème (e.g., PLSA, PLSM), il est possible de capturer différentes informations dans les thèmes. En particulier, le modèle PLSM permet à la fois d'extraire des thèmes-motifs contenant une information temporelle forte, et de déterminer quand ces motifs apparaissent.

Les thèmes, extraits de manière totalement non supervisée par les approches

présentées, permettent d'avoir un résumé extrêmement concis des activités présentes dans une scène. Au delà de la compréhension de scènes, ce chapitre a aussi présenté comment un modèle tel que PLSM peut être utilisé avec succès pour le comptage d'événements fréquents ou la détection d'activités anormales.

De part la qualité des résultats obtenus et leurs large domaine d'application, les modèles à thèmes ont un avenir certain dans le domaine de la reconnaissance d'activité dans les vidéos et documents multimédia.

Acknowledgements

The authors gratefully acknowledge the financial support from the Swiss National Science Foundation (Project : FNS-198,HAI) `www.snf.ch/E` and from the 7th framework program of the European Union project VANAHEIM (248907) `www.vanaheim-project.eu` under which this work was done.

Références

[BNJ03] D. M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Machine Learning Research*, (3) :993–1022, 2003.

[FGP08] K. Farrahi and D. Gatica-Perez. What did you do today ? Discovering daily routines from large-scale mobile data. In *ACM Int. Conf. on Multimedia (ACM MM)*, Vancouver, 2008.

[HFS08] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *Ubiquitous computing (UbiComp)*, pages 10–19, Seoul, Korea, 2008.

[Hof01] T. Hofmann. Unsupervised learning by probability latent semantic analysis. *Machine Learning*, 42 :177–196, 2001.

[NWL08] J-C. Niebles, H. Wang, and F.-F. Li. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79(3) :299–318, 2008.

[VEO10] J. Varadarajan, R. Emonet, and J.-M. Odobez. Probabilistic latent sequential motifs : Discovering temporal activity patterns in video scenes. In *Procedings of the British Machine Vision Conference (BMVC)*, Aberystwyth, 2010.

[VO09] J. Varadarajan and J.-M. Odobez. Topic models for scene analysis and abnormality detection. In *ICCV-12th International Workshop on Visual Surveillance (VS)*, 2009.

[WMG09] X. Wang, X. Ma, and E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 31(3) :539–555, 2009.