# Interprétation Automatique de séquences vidéos

François Brémond

INRIA Sophia Antipolis, Projet ORION
2004 route des lucioles - BP 93 FR-06902 Sophia Antipolis
francois.bremond@sophia.inria.fr
http://www-sop.inria.fr/orion/personnel/Francois.Bremond/

**Résumé** : Ces travaux présentent une approche d'interprétation de séquences vidéos pour la génération automatique d'annotations. Cette approche permet de détecter et de suivre des objets mobiles (e.g. personnes) dans des vidéos et de reconnaître des scénarios d'intérêt. L'interprétation vidéo met en jeu des techniques de vision par ordinateur, de vision cognitive, de représentation des connaissances et de techniques d'apprentissage. Cette approche a montré des résultats encourageants dans de nombreux domaines applicatifs tels que la surveillance de supermarchés, d'autoroutes, de métros, de trains, du tarmac d'aéroports et d'agences bancaires. Des limitations demeurent : hypothèses d'utilisation restrictives, robustesse des algorithmes de vision (segmentation, classification, suivi) et acquisition fastidieuse des modèles de scénario. Cet exposé montrera les performances et limitations de l'interprétation vidéo pour la génération automatique d'annotations associées aux vidéos. Cet exposé présentera les nouvelles tendances dans le domaine (e.g. utilisation d'ontologies) pour structurer la connaissance nécessaire à l'obtention de solutions opérationnelles.

**Mots-clés** : Interprétation d'images, reconnaissance de forme, reconnaissance de scénario, séquence d'images

## 1 Interprétation Vidéo

L'interprétation vidéo a pour objectif de générer automatiquement en temps réel des alarmes et en différé des annotations associées aux vidéos. Plus généralement, l'interprétation automatique d'images est une problématique difficile qui est la base de nombreux travaux en vision et aussi en intelligence artificielle. La difficulté dépend de la nature des entités à reconnaître et du type d'interprétation recherchée. Il est plus simple de reconnaître des objets statiques et rigides en environnement manufacturé, que des comportements dynamiques de plusieurs objets non-rigides en environnement naturel. La difficulté dépend également du type d'interprétation recherchée. Le problème peut être soit, simplement, d'étiqueter une entité bien déterminée que l'on peut mettre directement

en correspondance avec des modèles, soit de détecter les entités, de les étiqueter et de vérifier leur cohérence (spatiale, temporelle, structurelle, etc). Les résultats de l'interprétation peuvent être la reconnaissance d'objets physiques, d'événements, de situations ou de scénarios. L'interprétation de séquences d'images a pour objectif, pour ce qui nous concerne, de donner un sens à une scène décrivant des activités humaines, à partir d'images fournies par une caméra couleur, monoculaire et fixe. Cette interprétation de scène repose, en général, sur la coopération d'un module de traitement d'images, d'un module de suivi des objets mobiles et d'un module de reconnaissance du comportement des objets mobiles qui s'appuient sur une base de contexte. Il s'agit, pour le module de traitement d'images, de détecter les régions mobiles sur la séquence d'images. Le module de suivi associe les régions détectées afin de former et de suivre les objets mobiles. La tâche du module de reconnaissance des comportements consiste, grâce à des techniques d'intelligence artificielle, à identifier les objets suivis et à reconnaître leur comportement comme constitutif d'un ou plusieurs scénarios prédéfinis. Un point important dans l'interprétation vidéo est ainsi la représentation et la reconnaissance de scénarios.

## 2 Représentation des Scénarios

Un formalisme de représentation permet de définir un scénario comme composé d'états ou d'événements. Un état est une propriété spatio-temporelle définie à un instant donné ou sur un intervalle de temps alors qu'un événement est composé d'un ou plusieurs changements d'états entre au moins deux états successifs ou sur un intervalle de temps. De plus, un scénario peut-être primitif (changement d'état simple) ou composé (combinaison d'états et/ou d'événements) (cf. tableau 1). Il est constitué de trois éléments : les objets physiques contiennent une liste de personnes et d'objets réels intervenant dans le scénario, les composants contiennent une liste d'états et d'événements présents dans le scénario, et les contraintes contiennent une liste de relations entre les objets physiques et les événements. Les objets physiques sont des objets mobiles ou contextuels. Les objets mobiles sont généralement une personne, un véhicule ou un groupe de personnes. Les objets contextuels sont des zones prédéfinies (e.g. zone d'entrée, local sensible) ou du mobilier (e.g. guichets, chaises).

| | |
|---|---|
| **événement_composé** | Attaque_de_banque_avec_1personne |
| **objets_physiques :** | ( (p : personne), (z1 : derrière_guichet), (z2 : salle_sensible), (g : porte_salle_sensible) ) |
| **composants :** | ( c1 : événement_primitif changement_de_zone(p, z1, z2) ) |
| **contraintes :** | (g est ouverte) |

TAB. 1 – Événement composé avec une personne utilisant la primitive changement_de_zone : la personne p se déplace de la zone z1 vers la zone z2 et la porte est ouverte.

# 3 Algorithme de Reconnaissance de Scénarios

L'algorithme temps-réel de reconnaissance comporte plusieurs étapes :

- Dans un premier temps, l'algorithme reconnaît itérativement les états primitifs en sélectionnant un ensemble d'objets physiques et en vérifiant les contraintes atemporelles correspondantes jusqu'à ce que toutes les combinaisons d'objets physiques aient été testées.
- Ensuite, l'algorithme reconnaît les événements primitifs. Un événement primitif a deux états primitifs comme composants. Lors de la reconnaissance des états primitifs, des instances (événements partiellement reconnus) sont construites pour chaque événement se terminant par l'état primitif reconnu. Ces instances contiennent la liste des objets physiques et le dernier composant relatifs à l'état primitif reconnu. L'algorithme recherche ensuite si le premier composant de l'événement primitif correspond à un des états primitifs reconnus dans le passé. Si les deux composants vérifient les contraintes définies dans le modèle, alors l'événement primitif est reconnu.
- Enfin, l'algorithme reconnaît les états et les événements composés. La reconnaissance de ces états et événements implique une recherche exhaustive de toutes les combinaisons possibles des objets physiques et des composants. Pour limiter une explosion combinatoire, tous les états et événements composés sont découpés en des états et des événements contenant au plus deux composants pendant une phase de pré-compilation. La reconnaissance des états et des événements composés est ainsi ramenée à celle des événements primitifs.

# 4 Problèmes en Génération Automatique d'Annotations de Vidéos

L'interprétation vidéo a montré des résultats encourageants dans de nombreux domaines applicatifs tels que la surveillance de supermarchés, d'autoroutes, de métros, de trains, du tarmac d'aéroports, d'agences bancaires et plus récemment pour la surveillance médicale de personnes âgées ou d'enfants hospitalisés. Des limitations demeurent : hypothèses d'utilisation restrictives (caméras fixes et connaissances a priori sur la scène observée), robustesse des algorithmes de vision (segmentation, classification, suivi) et acquisition fastidieuse des modèles de scénario. Pour pallier au problème de l'acquisition des connaissances (modèles de scénario), deux pistes sont actuellement envisagées. Premièrement, l'utilisation d'interfaces homme/machine permet à l'expert de définir ces scénarios en étant guidé par une ontologie du domaine d'application composée d'états et d'événements vidéos. Deuxièmement, des techniques d'apprentissage permettent d'apprendre automatiquement des modèles de scénarios récurrents se déroulant dans la scène.