Traitement intégré d'informations vidéo pour la vidéosurveillance : le projet CAnADA .

Jacques BOONAERT[1], Lyès HAMOUDI[1]

[1]ARMINES-EMD, Département Informatique & Automatique, Ecole des Mines de Douai, 941 rue Charles Bourseul B.P 10838, 59508 DOUAI Cedex

boonaert@ensm-douai.fr, hamoudi@ensm-douai.fr

Résumé – Cet article est une description du projet de vidéosurveillance intelligente CAnADA. Après avoir rappelé les objectifs de celui-ci et détaillé la constitution du consortium, l'accent est mis sur les différents problèmes qu'il est nécessaire de résoudre pour aboutir à un système fonctionnel dans un contexte opérationnel le plus large possible. L'état de l'art fourni ici permet de mettre en perspective les apports attendus du projet. A titre d'illustration de certaines de préoccupations actuelles, nous détaillons un des algorithmes de suivi développé par l'équipe ARMINES-EMD. La dernière partie est l'occasion de dresser un premier bilan des actions menées et de dégager des perspectives à court terme.

Abstract – This paper is a description of the CAnADA video surveillance project. After briefly recalling the main objectives of this project and giving the detail of the involved teams, we focus on the problems we have to solve to get a system targeted to be effective in a wide spectrum of applications. The "state of the art" is provided to define what can be expected from CanADA, with respect to other projects sharing the same kind of objectives. As an illustration of some of the points that are addressed at these times, we give information concerning a tracking process developed by the ARMINES-EMD team. The last part of this paper allows us to make a first assessment of the project and to give indications related to the further works.

1. Description du projet CAnADA

1.1 But du projet

L'objectif du projet CAnADA (Comportements Anormaux : Analyse, Détection, Alerte), sélectionné par l' ANR dans le cadre de l'appel à projet CSOSG de 2006, est de fournir un ensemble d'outils permettant, principalement à partir d'informations issues d'un réseau de caméras vidéo, de caractériser, voire de prévoir, l'occurrence de comportements anormaux ou « à risque » affectant une ou plusieurs personnes fréquentant un lieu accueillant du public. La situation observée ayant été caractérisée, il s'agit ensuite d'acheminer l'information pertinente aux entités les plus à même de ramener celle-ci à un niveau normal en utilisant les canaux de retour d'information les plus appropriés (affichage, lignes spécifiques, etc.). Il s'agit donc, *in fine*, de proposer une solution « intégrée » pour la gestion de certains types d'alertes.

Deux aspects font l'objet d'attentions particulières dans le cadre de ce projet, à savoir le développement et la mise en œuvre de méthodes destinées à « donner du sens » à des flots de données tels que ceux issus de séquences d'images, ainsi que la capacité à manipuler efficacement la masse de données que constitue ce type d'informations.

Un autre thème central à ce projet réside dans la définition de modèles comportementaux qui seront exploités dans le cadre de l'évaluation de la menace et dans la détermination des moyens de communication (« canaux de retour ») à mobiliser pour tenter de gérer cette dernière.

1.2 Les équipes mobilisées.

Le consortium est dimensionné en fonction du très large spectre des problématiques couvertes par ce type de projet de vidéosurveillance intelligente. Ses compétences se rapportent aux domaines scientifiques, juridiques et industriels. Sur le plan académique, le projet tire profit des complémentarités existantes entre les différents laboratoires impliqués, c'est à dire le LIRIS (pour ce qui touche aux aspects « analyse d'objets en mouvement », « gestion des occultations », « reconnaissance de visages » et « indexation de données vidéo »), le LIFL - TÉLÉCOM LILLE 1 (pour les points relevant des problématiques traitées par l'équipe FoxMiire, dont en particulier la fouille de données complexes et multimédia, l'analyse des situations à un niveau « sémantique »), ARMINES-EMD (problème de gestion des flux vidéos multiples, suivi de trajectoires multiples en temps-réel, analyse « bas-niveau » des séquences de mouvements et classification de données évolutives), l'URECA (pour l'interprétation des

comportements individuels et collectifs), l'IREENAT (pour l'analyse des problèmes juridiques et de l'impact sociétal de la mise en œuvre de tels dispositifs). Les partenaires industriels, YOUG'S et Thales, constituent une interface de choix avec les prescripteurs potentiels des retombées industrielles du projet, tout en participant à l'industrialisation future des produits correspondants, en post-projet.

Dans ce qui va suivre, nous allons décrire plus en détail l'approche développée au sein de CAnADA, en la mettant en perspective avec « l'état de l'art » relatif à certains des principaux verrous identifiés ainsi qu'avec d'autres projets partageant des objectifs comparables. Par la suite, nous détaillerons certains travaux en cours de l'équipe ARMINES-EMD, représentatifs des activités en cours, pour ensuite indiquer les perspectives à court / moyen terme. La conclusion sera l'occasion d'un bref premier bilan sur ce projet.

1.3 Aspects traités dans le cadre du projet

En termes généraux, les informations auxquelles nous nous intéressons se ramènent pour partie aux déplacements et aux actions entreprises par une ou plusieurs personnes visibles dans les champs des caméras du réseau de vidéosurveillance. Ceci nécessite de prendre en compte des aspects liés à l'estimation des trajectoires, mais aussi à l'extraction de données liées à la « posture », à la « démarche » et à la reconnaissance des gestes effectués. Une des contraintes fortes pesant sur ces traitements est la nécessité d'effectuer ces derniers en un temps compatible avec les exigences des applications envisagées. Les méthodes développées à tous les niveaux de ce projet se doivent donc d'être performantes à la fois sur le plan de leur « sélectivité » (taux de reconnaissance en particulier) mais aussi sur celui de leur « efficacité » en termes algorithmiques.

De façon précise, les principales données prises en considération, pour chaque personne observée, sont :

- Sa trajectoire, suite temporelle de positions spatiales après correction des occultations,
- L'« activité », qui peut se concevoir ici comme la suite des actions « élémentaires » (avec les durées correspondantes) menées par une personne sur un intervalle de temps donné.

A terme, ces actions pourront se rapporter aussi aux interactions avec les autres individus, l'environnement et les objets de l'environnement. La détermination des actions élémentaires mobilise des techniques de classification permettant d'associer celles-ci à une séquence d'images. La capacité d'adaptation du système à son contexte d'exploitation implique la faculté à détecter l'occurrence d'actions élémentaires non répertoriées afin d'enrichir dynamiquement l'ensemble des prototypes connus. Ceci nous place *de facto* dans un contexte semi-supervisé. La figure ci-dessous retrace la succession des traitements correspondants :

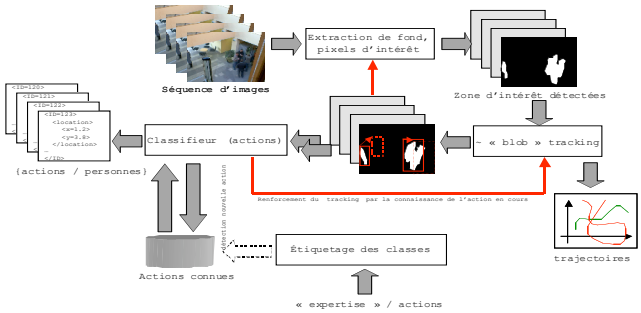

FIG. 1 : séquence de traitements pour l'extraction des informations de base.

L'objectif de cette première chaîne de traitements est d'aboutir à une description des scènes observées dotée d'un degré d'abstraction suffisamment élevé pour autoriser l'utilisation d'algorithmes liés aux tâches de classification et d'apprentissage en ligne ainsi que de « décision », sur lesquels repose le système de surveillance.

L'aspect « distribué » du système, tant sur le plan spatial (exploitation d'un réseau de caméras) que temporel (possibilité de recouper des informations correspondant à des instants d'observation différents) impose, là encore, le développement de méthodes adaptées. Ainsi, en plus de l'extraction d'informations à caractères « dynamiques » liées à l'attitude des personnes observées, l'extraction de données de nature « descriptives » doit être traitée afin de pouvoir associer un même individu à différentes zones surveillées, que ce soit de façon simultanée (présence dans le champ de plusieurs caméras) ou à différents instants (suivi du parcours sur l'intégralité du site surveillé, repérage de passages répétés sur différentes périodes, etc.). Il est clair que la définition d'un ensemble de « vecteurs formes » à la fois suffisamment robustes et spécifiques est un point délicat, compte tenu de la nature antagoniste de ces exigences.

Un autre point d'achoppement de ce projet concerne la notion de « normalité ». Celle-ci est traitée en fonction du contexte et par l'intermédiaire de « modèles comportementaux », indispensables à l'analyse des données de haut niveau obtenues à partir de la vidéo. La démarche consiste alors à confronter les résultats de ces différents modèles aux observations délivrées par le système afin de raisonner sur des « résidus comportementaux », dans le but de caractériser les comportements perçus, qu'ils soient individuels ou collectifs.

La classification effectuée doit pouvoir répondre à la question de la « normalité » de la situation compte tenu du

contexte et des éventuelles dynamiques de groupe mises en évidence : Il nous faut pouvoir répondre à des questions du type « S'agit-il d'un comportement normal compte-tenu du contexte ou de l'attitude des autres personnes ? ». En d'autres termes, il s'agira à partir de l'ensemble des informations disponibles de décider si les activités d'une ou plusieurs personnes sont signifiantes d'une action (qu'elle soit en cours ou « probable » à court ou moyen terme) propre à compromettre la sécurité des personnes et des biens sur le site surveillé. A ce stade se pose une difficulté : il peut être plus simple de définir une situation normale qu'une situation anormale, qu'on cherche « par construction » à éviter ! Faute d'informations suffisantes *a priori*, il faut donc mettre en œuvre des méthodes qui permettront d'agréger en classes homogènes les ensembles de comportements (qu'ils soient individuels ou collectifs) ayant été détectés comme « non-normaux », ceci afin de pouvoir affiner la caractérisation de la situation. Il faudra par conséquent que le système intègre, là aussi, des fonctions d'apprentissage à partir de données par nature évolutives, appliquées à un système (le lieu surveillé) dont le « mode de fonctionnement » (état normal, état d'alerte) est sujet à des variations qui peuvent être brusques.

La figure ci-dessous représente une structure possible de la chaîne de traitement associée aux fonctions de classification et d'apprentissage en ligne :

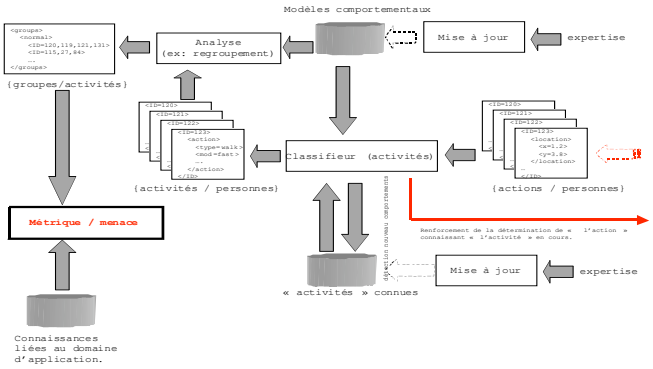

FIG. 2 : séquence de traitements pour les tâches de classification et d'apprentissage en ligne.

Afin de gérer en partie l'alerte, le système doit être capable de diffuser les informations pertinentes aux acteurs les plus à mêmes de ramener la situation à la normale. Le fait qu'une alerte puisse être évaluée implique qu'une « métrique » puisse être associée à la situation observée. Celle-ci dépendra nécessairement très largement du contexte et mobilisera les connaissances et le savoir-faire des spécialistes de la sécurité associés à CAnADA. Un vaste travail de formalisation associé à un ensemble de cadres applicatifs clairement identifiés est donc intégré au projet.

Même si cet article décrit davantage les aspects liés au traitement des séquences vidéo, il convient de noter que les apports des techniques associées à l'indexation et à la fouille de données complexes et multimédia sont fondamentaux et transversaux à la plupart des travaux. En effet, en plus de données recueillies sur le terrain en temps réel, l'évaluation de la menace va potentiellement mobiliser les connaissances antérieurement acquises (parcours et activités précédentes sur de longues périodes, caractéristiques d'individus « à risque » précédemment signalés, types de comportements anormaux déterminés lors de phase d'apprentissage, données associées au contexte, etc.) afin d'y effectuer les recoupements nécessaires.

Les objectifs et les principaux domaines couverts par CAnADA ayant été décrits, nous donnons dans la section suivante un aperçu de quelques projets dont les objectifs sont comparables ainsi qu'un bref état de l'art associé au traitement des séquences vidéo, ceci afin de pouvoir mieux cerner nos spécificités.

1.4 Contexte scientifique associé au projet.

1.4.1 D'autres projets de vidéosurveillance intelligente.

Compte tenu du potentiel de l'approche, les projets s'articulant autour de la vidéosurveillance intelligente sont relativement nombreux. Le descriptif ci-dessous a pour objet de citer quelques éléments jugés représentatifs dans le domaine du suivi de personne et n'a aucunement prétention à l'exhaustivité :

- VSAM : *Video Surveillance and Monitoring* [22]. Ce projet, supervisé par l'Université Carnegie Mellon et l'Institut Sarnoff, sous financement DARPA, a étudié un ensemble de techniques d'analyse de vidéo pour des applications de surveillance d'environnement urbain (et de champ de bataille), permettant de signaler à un opérateur les incidents détectés par un vaste ensemble de capteurs.
- ADVISOR : *Annotated Digital Video for Surveillance and Optimised Retrieval* [9]. Ce projet européen (IST-1999-11287) a étudié la détection d'incidents et l'annotation d'enregistrements vidéo pour des problèmes de surveillance, avec une application à la surveillance du métro.
- MIT CSAIL : [8]. Dans le cadre du projet VSAM, l'équipe Vision du MIT a particulièrement développé des techniques pour gérer de façon automatisée un ensemble de caméra fixes permettant de visualiser complètement une large zone de l'environnement.
- CAVIAR : *Context Aware Vision using Image-based Active Recognition* [7]. Ce projet européen (IST 2001 37540) étudie des techniques d'analyse fine des images afin d'améliorer les performances

des systèmes de surveillance, avec des applications à l'environnement urbain et à celui des activités commerciales.
- GMF4iTV : *Generic Media Framework for Interactive TV* [1]. Ce projet européen a développé des techniques de suivi d'objets dans des séquences vidéo afin de créer des liens hyper-vidéo dans des programmes de télévision interactive, ces liens permettant d'associer des méta-données à des objets identifiés dans les images.

En dépit de la grande diversité des domaines d'application abordés, la détection des comportements anormaux associée à la prise en compte d'informations relatives au contexte ainsi qu'au « mode de fonctionnement » du site surveillé, reste relativement peu traitée.

Par ailleurs, la partie « validation » est un point là aussi relativement peu développé, ne serait-ce que par la difficulté d'accéder à un corpus de données suffisamment volumineux, représentatif et « correctement étiqueté », tout en respectant les dispositions légales en vigueur. La constitution de tels corpus de données fait partie des objectifs du projet Canada.

Ce dernier se positionne donc avec les particularités de vouloir offrir une approche « intégrée » allant de l'observation des comportements jusqu'à une première gestion de l'alerte, tout en affichant des capacités d'adaptation au contexte (liées au développement d'outils de classification et d'apprentissage en ligne) lui permettant de se déployer dans différents domaines applicatifs. Les questions liées au droit et à « l'acceptabilité » des solutions proposées sont prises en compte dès la genèse du projet (en particulier pour ce qui touche à la constitution des corpus de données, comme indiqué plus haut).

1.4.2 Etat de l'art associé au problème de suivi des personnes

S'agissant de l'un des domaines imposant le plus de « verrous », nous détaillerons plus particulièrement ici l'état de l'art relatif à l'extraction des déplacements à partir de la vidéo, connu aussi sous le terme de « suivi de personnes », qui est particulièrement riche. L'intérêt pour ces problèmes a été largement suscité par les multiples applications de la vidéosurveillance. La plupart des références citées ci-après ne constituent qu'un échantillon de travaux représentatifs des approches actuelles dans le domaine. Précisons que nous nous plaçons ici sous l'hypothèse de caméras fixes. Une approche commune en la matière consiste en une analyse fine de la première image d'une séquence vidéo pour identifier des zones potentielles où pourraient se trouver une ou plusieurs personnes. Un suivi de ces zones dans les images suivantes de la séquence est ensuite effectué. On peut distinguer trois catégories de modélisations permettant l'identification et le suivi de ces zones :

- la modélisation par région [17], qui identifie des régions homogènes par la couleur ou la texture, puis assemble ces régions en une « personne »,
- les modèles 2D, qui essaient de représenter les contraintes de l'image d'une personne, soit comme zone intérieure d'un contour, soit comme un ensemble articulé de régions. En raison de la forme des entités visuelles traitées, les contours sont souvent représentés par des courbes de type b-spline ou snake-spline. L'évolution au cours du temps des paramètres associés à ces courbes permet par ailleurs une analyse du mouvement.
- les modèles 3D, qui utilisent une modélisation en volume de la personne pour retrouver la trace de ses mouvements dans la séquence vidéo.

Qu'il s'agisse de modèles 2D ou 3D, dès lors que la personne est modélisée par une collection d'objets (rigides ou non) articulés entre eux, le suivi effectué ne s'opère plus sur un point de référence unique (par exemple le centre de gravité d'un ensemble de pixels, encore appelé « blob ») mais sur un ensemble de « points d'intérêt » définis de sorte à permettre la reconstruction de la posture de la personne. La définition de ces points d'intérêts de façon entièrement automatique et sans utilisation de marqueurs visuels portés à dessein (tel que cela est pratiqué dans le cadre de certaines applications d'analyse du mouvement) est un problème en soi. Une solution, mise à profit dans le cadre de projet, peut par exemple être apportée par l'exploitation de l'information « couleur » [19], souvent caractéristique des visages et des mains, qui permet alors de définir des points d'ancrage pour l'adaptation d'un modèle sur une image. Dans le même ordre d'idée, l'utilisation de l'information texture ou de modèles spécifiques à certaines parties du corps peut s'avérer payante. On trouve par ailleurs des approches permettant de traiter la phase d'initialisation de ce type de modèle articulé à partir d'un ensemble suffisant de points d'intérêt [2]. D'autres problèmes sont à prendre en considération, comme les auto occultations et le port de certains vêtements (masquage de vastes parties du corps).

Comme indiqué, les différents algorithmes utilisés sont initiés par la détection de zones pertinentes dans les séquences d'images. Sous l'hypothèse de l'utilisation d'une (ou plusieurs) caméra(s) statique(s), les objets à détecter se distinguent par leurs mouvements par rapport à l'image de fond. Une approche relativement classique consiste à déterminer cette image de fond afin de pouvoir ultérieurement la supprimer dans les séquences [26]. Les traitements effectués peuvent ensuite être de type « blob tracking », où les pixels correspondant aux zones d'intérêt font l'objet d'opérations morphologiques de type dilation / érosion. Les pixels restant sont alors agrégés pour former les zones d'intérêt soumises aux traitements ultérieurs. D'autres auteurs pré traitent directement l'image (à l'aide par exemple de filtres à réponse impulsionnelle infinie) afin d'extraire des informations relatives au mouvement apparent des pixels [14].

Des filtres de Kalman [5,25] sont souvent utilisés pour faire l'estimation de l'évolution temporelle des paramètres du modèle le long de la séquence. Un autre intérêt de ce type de méthode est de fournir de façon « naturelle » une zone de présence probable (R.O.I pour « *Region Of Interest* ») de l'entité poursuivie (supposée être la personne) dans la séquence. Compte tenu des traitements complexes qu'implique la manipulation d'images, ceci permet de diminuer de façon drastique les temps de calculs en restreignant ces derniers aux zones d'intérêt détectées.

Précisons que diverses techniques de modélisation probabiliste comme les Pseudo Modèles de Markov Cachés en 2D [24] ou le filtrage particulaire [3] sont employées dans ce contexte d'analyse des trajectoires.

Les conditions d'exploitation des méthodes d'extraction de parcours vont varier suivant l'infrastructure utilisée et/ou les comportements typiques des personnes situées en ces lieux : les déplacements observés ne seront *a priori* pas les mêmes dans une allée passante et dans un lieu d'attraction (étale, animation, etc.). Les algorithmes proposés sont donc à distinguer suivant qu'ils permettent :

- le suivi d'une ou de plusieurs personnes [21],
- le traitement des occultations par des objets ou des rencontres entre des personnes,
- l'utilisation d'une ou de plusieurs caméras [18], ce qui pose en particulier le problème d'identifier les zones en recouvrement visibles sur plusieurs caméras en même temps [13]. Des techniques de calibration du réseau de caméra peuvent cependant être mises en œuvre de façon suffisamment souple pour être exploitées dans le contexte du projet [6]. Les zones de recouvrement peuvent par ailleurs être exploitées en profitant de la mise en correspondance spatiale et temporelle qu'elles autorisent, permettant la mise en œuvre de techniques issues de la stéréovision.
- l'utilisation de caméras mobiles, calibrées ou non,
- la détection et analyse de mouvements de foule [3],
- la satisfaction de contraintes exprimées en terme de temps de calcul (temps réel, traitement simultané de plusieurs flux vidéo) [16].

L'application d'une procédure de classification permet, au travers de l'analyse des parcours et des postures des personnes détectées, d'associer un type d'action [14], voire un comportement à ces dernières. Les applications envisageables sont alors (par exemple) la détection d'incidents dans le métro [4] ou la reconnaissance de personnes par la démarche [20]. En matière de reconnaissance des actions, les approches proposées peuvent se diviser en deux classes. Il est ainsi possible d'effectuer la reconnaissance de l'action à partir de l'analyse de la trajectoire de chacun des « objets » rigides associés au corps de la personne. L'inconvénient d'une telle démarche est la difficulté de distinguer deux activités pourtant différentes qui se caractériseraient par des ensembles de trajectoires comparables. L'autre possibilité consiste à baser la reconnaissance sur le changement d'apparence de la personne dans la séquence d'images (au travers d'une modélisation adéquate des contours ou des régions, par exemple). Ces changements sont ensuite comparés à des prototypes préalablement mémorisés. Cette démarche nécessite une phase de normalisation de la séquence d'images, du fait de sa sensibilité au facteur d'échelle, rotation et translation. Par ailleurs, les phénomènes d'occultation peuvent rendre difficile la mise en correspondance. Des approches mixtes modélisent le changement d'apparence par le biais de l'exploitation de modèle 3D, permettant de gérer efficacement ces phénomènes d'occultation. Ce type de démarche implique de prendre en considération le « couplage » avec les tâches de tracking et d'estimation de la trajectoire. Malgré les nombreux travaux réalisés, les performances des techniques de suivi de personnes ne sont pas encore idéales. En particulier, des progrès sont encore à réaliser sur le problème des occultations partielles et sur l'appariement des occurrences d'une même personne sur une autre séquence ou après une occultation totale. Par ailleurs, les solutions seront déployées dans des lieux connaissant potentiellement une très forte fréquentation (point de vente, lieux de fort passage), ce qui implique le développement d'algorithmes particulièrement efficaces (en terme de temps de calcul) afin que la mise en œuvre puisse se faire sur un matériel le plus standard possible afin de ne pas impliquer un coût trop élevé.

A titre d'illustration, la section suivante présente une méthode de suivi de personne s'inscrivant dans le schéma global des traitements appliqués aux séquences vidéos et reflétant les premiers résultats issus de ce projet dans ce domaine.

2. Un exemple parmi les premiers résultats : détection et suivi de visage dans une séquence vidéo.

2.1 Introduction.

Cette section présente une technique pour la détection et le suivi (*tracking*) automatique d'un visage dans une séquence vidéo, actuellement développée au sein du Département Informatique et Automatique de l'Ecole des Mines de Douai. Les raisons ayant conduit notre équipe à traiter cet aspect sont de deux ordres :

- en premier lieu, comme indiqué dans la section relative à l'état de l'art, la détection d'éléments spécifiques à la personne humaine (dont le visage est en exemple pertinent) s'avère particulièrement utile pour initialiser les différents modèles (2D ou 3D) qui seront mis à contribution pour effectuer les analyses ultérieures (détermination des postures, reconnaissance de geste, etc). Ajoutons que les premières étapes du traitement mis en œuvre peuvent être exploitées pour présélectionner des pixels associés à d'autres parties du corps (les mains en particulier).
- En second lieu, la constitution d'un corpus de données représentatif et exploitable sans

contraintes trop lourde à gérer en matière de droit va nécessiter « l'anonymisation » des séquences contenant l'image de tierces personnes n'étant pas directement impliquées dans les équipes de recherche associées au projet. Une possibilité acceptable d'anonymisation consiste alors à coder automatiquement les zones des images correspondant aux visages de ces personnes, d'où le besoin de disposer d'une technique à la fois fiable et efficace.

Conformément au schéma de traitement global décrit précédemment, la méthode retenue est basée sur un processus en plusieurs étapes qui détecte dans un premier temps dans une image couleur les régions qui sont susceptibles de contenir de la peau et extrait ensuite de ces régions les informations qui indiquent l'endroit dans lequel se trouverait le visage.

Plus précisément, l'espace colorimétrique RGB initial est transformé de manière à extraire des informations de couleur et de texture, qui s'avèrent en pratique plus efficaces pour sélectionner les pixels pouvant représenter de la peau. De façon désormais classique, ces pixels sont ensuite « structurés » par le biais d'opérations morphologiques. La détection d'un visage découle ensuite de l'application de critères géométriques. En ce qui concerne le suivi, l'application d'un filtre de Kalman reposant sur une modélisation simple de la cinématique apparente de « l'objet » suivi sur un cours intervalle de temps donne ici de bons résultats.

2.2 Principe de détection.

La toute première étape, qui consiste en la sélection de « pixels d'intérêt » est caractéristique d'approches basées sur l'exploitation « d'invariants » (*features invariant*). L'image d'entrée est au format RGB. Celle-ci est transformée en valeurs log-opponent (IRgBy)[27], qui permet par la suite le calcul de composantes *Texture*, *Teinte*, et *Saturation*. La conversion du format RGB au format log-opponent (IRgBy) est effectuée selon les formules suivantes :

$$I = L(G) \quad (1)$$

$$Rg = L(R) - L(G) \quad (2)$$

$$By = L(B) - (L(G) + L(R)) / 2 \quad (3)$$

La fonction $L(x)$ est, quant à elle, définie par :

$$L(x) = 105.\log(x+2) \quad (4)$$

Les matrices Rg et By sont alors filtrées avec un filtre médian carré de côté *4.échelle*. La valeur *échelle* est calculée comme étant l'entier le plus proche de la quantité (*hauteur*+*largeur*)/320, où *hauteur* et *largeur* se rapportent aux dimensions de l'image traitée. L'image *Texture* est employée pour trouver des régions de faibles valeurs de texture. En effet, la peau, telle qu'elle apparaît dans les images, tend à avoir une texture très lissée. Cette composante *Texture* est produite à partir de la matrice *I* par l'intermédiaire des étapes suivantes :

1- Filtrer la matrice I par un filtre médian de côté *8.échelle*.
2- Soustraire cette matrice filtrée de la matrice originale *I*.
3- Prendre la valeur absolue de cette différence et filtrer par un filtre médian de côté *12.échelle*.

En complément de la texture, les images *Teinte* et *Saturation* sont utilisées pour discriminer les régions pouvant correspondre à de la peau. La conversion de log-opponent vers *Teinte* se fait par [27] :

$$Teinte = [atan2(Rg,By).(180/pi)] \quad (5)$$

La conversion de log-opponent vers *Saturation* se fait par [27]:

$$Saturation = (Rg^2+By^2)^{1/2} \quad (6)$$

La sélection des pixels « candidats » s'opère alors dans le nouvel espace (*Texture*, *Teinte,Saturation*) qui s'avère bien adapté au problème. Les méthodes de classification utilisables sont très nombreuses et leur choix doit être guidé par la recherche d'un bon compromis entre performances et complexité. A titre d'exemple, l'assimilation de la classe « pixel de peau » à un parallélépipède de (*Texture*, *Teinte,Saturation*) défini par :

$1 < Texture < 22$
$30 < Teinte < 180$
$10 < Saturation < 40$

donne déjà des résultats intéressants, illustrés par la suite de figures ci-dessous :

FIG. 3.a : Image originale

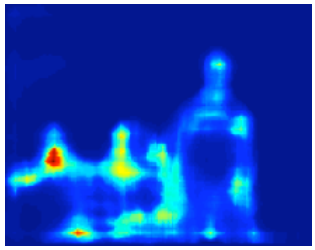

FIG. 3.b : Image texture

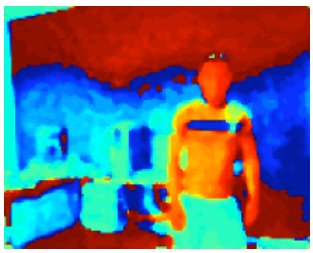

FIG. 3.c : Image teinte

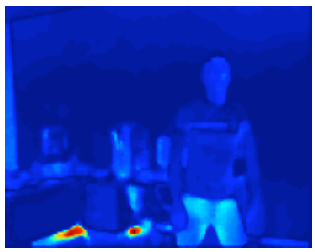

FIG. 3.d : Image saturation

FIG. 3.e : Pixels « peau »

FIG. 3.f : morphologies

Comme le montre la dernière figure de la série ci-avant, des opérations de morphologie mathématique sont appliquées à l'image binaire (dilatation, ouverture, remplissage) pour éliminer le bruit et constituer des régions homogènes.

Dans l'image binaire ainsi obtenue se trouvent plusieurs régions « candidates ». Parmi celles-ci, certaines correspondent effectivement aux parties du corps contenant de la peau (visage, mains,…) tandis que d'autres sont de « faux positifs » qu'il convient d'éliminer. Un filtrage supplémentaire est alors effectué en appliquant un jeu de critères géométriques aux régions détectées. Au nombre des critères applicables, nous pouvons noter la taille, la symétrie et l'excentricité.

A l'issue de ce nouveau filtrage, ne restent potentiellement que des « blobs » pouvant correspondre au visage d'une personne. Là encore, la discrimination repose sur l'application d'un ensemble de contraintes représentant les connaissances *a priori* que nous possédons sur l'aspect d'un visage (ce qui est, cette fois-ci, une caractéristique des méthodes « basées connaissance »). Dans le cas présent, cette opération s'appuie sur une phase de détection des yeux et / ou des narines exploitant une binarisation des niveaux de gris des pixels des régions candidates.

La série de figures ci-dessous illustre les résultats obtenus par ce procédé :

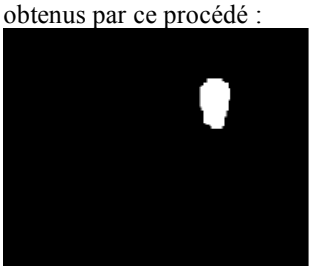

FIG. 4.a critères géométriques

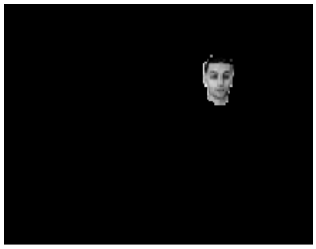

FIG. 4.b : Niveaux de gris

FIG. 4.c : Région binarisée

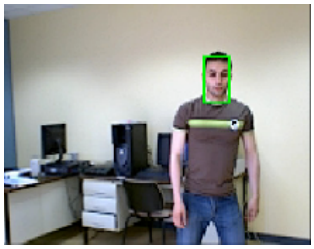

FIG. 4.d : Visage détecté

Pour le Suivi (*tracking*), nous calculons les coordonnées *(x,y)* du centre du rectangle associée à la région associée à un visage et utilisons le formalisme du filtrage de Kalman, basé sur un modèle cinématique simple, pour estimer sa prochaine position ainsi que sa taille (à partir de la variance d'erreur d'estimation). Le processus de détection de visage précédemment décrit est alors appliqué à cette fenêtre « prédite ». Si le visage n'y est pas détecté, le processus est réinitialisé (la recherche est appliquée de nouveau sur l'image entière).

Les tests effectués jusqu'à présent l'ont été dans des environnements relativement bien maîtrisés en termes de conditions de prise de vue (bureaux, laboratoire) et d'apparence des personnes suivies et les résultats obtenus sont très encourageants. Cependant, il est clair qu'un vaste potentiel de progression demeure dans l'exploration de méthodes de classification en ligne bien plus performantes que celle donnée à titre d'exemple dans cet article.

2.3 Conclusion sur la méthode présentée.

Bien que la méthode de suivi présentée n'en soit pour l'heure qu'à ses préliminaires, elle est néanmoins représentative des processus à plusieurs étapes dont le développement est indispensable à l'accomplissement des objectifs du projet. Les développements en cours relativement à ce module de suivi portent sur l'implémentation de « classifieurs » évolutifs et adaptatifs, capables de réaliser une classification dynamique et un apprentissage en ligne. Les compétences de l'équipe en matière de diagnostic nous orientent prioritairement vers les *Support Vector Machines* (SVM). En effet, ci ces derniers ont été utilisés avec succès dans diverses applications de reconnaissance d'objets (visages, piétons, véhicules) et ont démontré leur efficacité, leur capacité à réaliser une classification dynamique n'a, semble-t-il, pas encore été exploitée complètement dans le domaine du suivi d'entités dans les séquences d'image.

3. Autres développements en cours.

Comme indiqué, les travaux décrits dans la section précédente ne concernent directement qu'ARMINES-EMD et correspondent aux premières « productions » issues d'une phase préalable majoritairement dédiée aux études bibliographiques. En complément de ce qui a été présenté, l'ensemble des partenaires a contribué aux travaux relatifs à la définition du corpus de validation. Une difficulté supplémentaire vient ici de la nécessité de respecter les recommandations de la CNIL dans la constitution de ce dernier, qui est une des raisons ayant motivé les éléments présentés en 2. L'IREENAT a tout naturellement pris une part très active dans la prise en compte de ces contraintes et fut le principal artisan d'un rapprochement avec la CNIL actuellement en cours.

Des approches complémentaires à celle présentée en matière de suivi de personnes constituent l'actualité des développements au sein du LIFL et du LIRIS, tandis que l'URECA contribue à la définition des modèles comportementaux qui seront exploités par la suite. Ces éléments autoriseront très prochainement le traitement de la classification des postures, conformément au schéma de la figure 1.

4. Conclusion.

Dans cet article, nous avons présenté les objectifs du projet CAnADA ainsi que les différents domaines dans lesquels celui-ci apportera sa contribution afin de déboucher sur un système efficace dans une large gamme de conditions opérationnelles. Bien que l'approche « intégrée » de ce projet implique une très vaste problématique, nous avons ici plus particulièrement insisté sur les aspects liés au traitement des séquences vidéo, qu'il s'agisse des problèmes liés au suivi des personnes ou de ceux associés aux différentes phases de classification (classification des « actions » et classification des « activités »). Comme nous l'avons précisé, il ne s'agit ici

que d'un sous-ensemble des « verrous » du projet (qui comprennent aussi, entre autres, des éléments liés à la fouille de masse de données multimédia, à « l'acceptabilité » du système et aux contraintes techniques et de coûts). Comme a permis de le préciser l'état de l'art fourni ici, les difficultés auxquelles nous sommes confrontées sont de différents ordres : il faut ainsi définir des algorithmes qui soient suffisamment performants tout en respectant des contraintes liées au temps de réponse tolérable pour les applications envisagées. Par ailleurs, les capacités d'adaptation en ligne sont aussi un des points clefs, ce qui nécessite l'emploi de techniques issues des toutes récentes production de la recherche en la matière. Les éléments généraux du projet ayant été présentés, nous avons détaillé un des algorithmes de suivi développé au sein de notre équipe, dans la mesure où celui-ci sera exploité dans le cadre de la constitution des corpus de test et de la validation et du fait de son caractère significatif des problèmes actuellement traités par les différentes équipes. Hormis la difficulté à intégrer les éléments liés au droit dans la constitution des corpus de validation définitifs, l'état d'avancement du projet CAnADA est compatible avec le calendrier initial. Une fois les solutions actuellement développées pour le suivi de personne arrivées à maturité, les travaux nous concernant à court terme toucheront principalement les processus liés à la classification et à l'apprentissage en ligne.

Références

[1] F. de Carvalho, G. Fernandez, B. Merialdo, A. Navarro et G. Thallinger. *Hyperlinked video with moving objects in digital television*. International Conference on Multimedia and Expo ICME 2005, Amsterdam, July 2005.

[2] B. Li, Q. Meng et H. Holstein. *Reconstruction of segmentally articulated structure in freeform movement with low density feature points*. Image and Vision Computing, Volume 22, Issue 10, September 2004 Page(s) 749-759.

[3] N. Checka, K.W. Wilson, M.R Siracusa et T. Darrel. *Multiple person and speaker activity tracking with a particle filter*. ICASSP May 2004.

[4] F. Cupillard, A. Avanzi, F. Bremond et M. Thonnat. *Video understanding for metro surveillance*. IEEE International Conference on Networking, Sensing and Control, March 2004.

[5] V. Girondel, A. Caplier et L. Bonnaud. *Real time tracking of multiple persons by Kalman filtering and face pursuit for multimedia applications*. 6th IEEE Southwest Symposium on Image Analysis and Interpretation, March 2004.

[6] Ch. Jaynes. *Multi-view calibration from planar motion trajectories*. Image and Vision Computing Volume 22, Issue 7, July 2004 Page(s) 535-550.

[7] P.M. Jorge, A.J. Abrantes et J.S. Marques. *Estimation of the Bayesian Network Architecture for Object Tracking in Video Sequences*. ICPR, Cambridge, August 2004.

[8] A. Rahimi, B. Dunagan et T. Darrell. *Tracking People with a Sparse Network of Bearing Sensors*. European Conference on Computer Vision (ECCV), 2004.

[9] N. Siebel et S. Maybank. *The ADVISOR Visual Surveillance System*. Proceedings of the ECCV 2004 workshop "Applications of Computer Vision".

[10] F. Souvannavong, B. Mérialdo et B. Huet. *Improved video content indexing by multiple latent semantic analysis*. CIVR'04, International Conference on Image and Video Retrieval, July 21-23, 2004, Dublin City University, Ireland.

[11] F. Souvannavong, L. Hohl, B. Mérialdo et B. Huet. *Using structure for video object retrieval*. CIVR'04, International Conference on Image and Video Retrieval, July 21-23, 2004, Dublin City University, Ireland.

[12] T.H Chen. *An automatic bi-directional passing-people counting method based on color image processing*. International Carnahan Conference on Security Technology, Oct. 2003

[13] S. Khan et M. Shah. *Consistent labeling of tracked objects in multiple cameras with overlapping fields of view*. IEEE Pattern Analysis and Machine Intelligence,Volume 25, Issue 10, Oct. 2003 Page(s):1355 – 1360.

[14] O. Masoud et N. Papanikolopoulos. *A method for human action recognition*. Image and Vision Computing Volume 21, Issue 8, August 2003 Page(s) 729-743

[15] W. Niu, L. Jiao, D. Han et Y.F. Wang. *Real-time multiperson tracking in video surveillance*. Fourth International Conference on Information, Communications and Signal Processing, Dec. 2003

[16] J. Ruiz-del-Solar, A. Shats et R. Verschae. *Real-time tracking of multiple persons*. 12th International Conference on Image Analysis and Processing, Sept. 2003 Page(s):109 - 114

[17] N. Siebel. *Design and Implementation of People Tracking Algorithms for Visual Surveillance Applications*. PhD thesis, Department of Computer Science, The University of Reading, Reading, UK, March 2003.

[18] Meng Howe Tan et S. Ranganath. *Multi-camera people tracking using Bayesian networks*. Fourth International Conference on Information, Communications and Signal Processing, Dec. 2003

[19] L. Bretzner, I. Laptev et T. Lindeberg. *Hand gesture recognition using multi-scale colour features, hierarchical model and particle filtering*. 5th IEEE International Conference on Automatic Face and Gesture Recognition, 2002.

[20] L. Lee et W.E.L Grimson. *Gait analysis for recognition and classification*. 5th IEEE International Conference on Automatic Face and Gesture Recognition, May 2002.

[21] Hung-Xin Zhao et Yea-Shuan Huang. *Real-time multiple-person tracking system*. International Conference on Pattern Recognition, Aug. 2002

[22] Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto et Hasegawa. *A System for Video Surveillance and Monitoring: VSAM Final Report*. Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May, 2000.

[23] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale et S. Shafer. *Multi-camera multi-person tracking for EasyLiving*. 3rd IEEE International Workshop on Visual Surveillance, July 2000

[24] Y. Ricquebourg et P. Bouthemy. *Real-time tracking of moving persons by exploiting spatio-temporal image slices*. IEEE Pattern Analysis and Machine Intelligence, Volume 22, Issue 8, Aug. 2000 Page(s):797 - 808

[25] G. Rigoll, S. Eickeler et S. Muller. *Person tracking in real-world scenarios using statistical methods*. 4th IEEE International Conference on Automatic Face and Gesture Recognition, March 2000 Page(s):342 - 347

[26] A. Lipton, H. Fujiyoshi, et R. Patil. *Moving Target Classification and Tracking from Real-time Video*. Proc. DARPA IU Workshop, Monterey, CA. 1998. Page(s). 129-136

[27] J.P. Kapur, EE499 Capstone Design Project Spring, 1997