Fusion d'informations sur des images sursegmentées : Une application à la compréhension de scènes routières

Philippe Xu, Franck Davoine, Thierry Denoeux, Jean-Baptiste Bordes

Fusion d'informations sur des images sursegmentées : Une application à la compréhension de scènes routières

Philippe Xu[1,2] Franck Davoine[2] Thierry Denœux[1] Jean-Baptiste Bordes[1]

[1]UMR CNRS 7253 Heudiasyc
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex France

[2]LIAMA, CNRS
Peking University
Pékin, R.P. Chine

philippe.xu@hds.utc.fr

Résumé

Le grand nombre de tâches qu'on peut attendre d'un système d'aide à la conduite implique de considérer de nombreuses classes d'objets pouvant être à proximité du véhicule. Afin de comprendre correctement la scène, toutes les sources d'information disponibles doivent être combinées. Dans cet article, un schéma original de fusion au niveau des segments d'une image sursegmentée et basé sur la théorie des fonctions de croyance est présenté. Le problème est posé comme celui d'une annotation d'image. L'approche sera tout d'abord appliquée à la détection du sol en utilisant trois capteurs différents. La flexibilité du cadre de fusion permet d'ajouter facilement de nouveaux capteurs mais aussi de nouvelles classes, elle sera démontrée en ajoutant les classes ciel et végétation. Ce travail est validé sur des données réelles de scènes routières en milieu urbain.

Mots clés

Fusion d'informations, compréhension de scènes routières, théorie des fonctions de croyance, véhicules intelligents.

Abstract

The large number of tasks one may expect from a driver assistance system leads to consider many object classes in the neighborhood of the so-called intelligent vehicle. In order to get a correct understanding of the driving scene, one has to fuse all sources of information that can be made available. In this paper, an original fusion framework working on segments of over-segmented images and based on the theory of belief functions is presented. The problem is posed as an image labeling one. It will first be applied to ground detection using three kinds of sensors. We will show how the fusion framework is flexible enough to include new sensors as well as new classes of objects, which will be shown by adding sky and vegetation classes afterward. The work was validated on real and publicly available urban driving scene data.

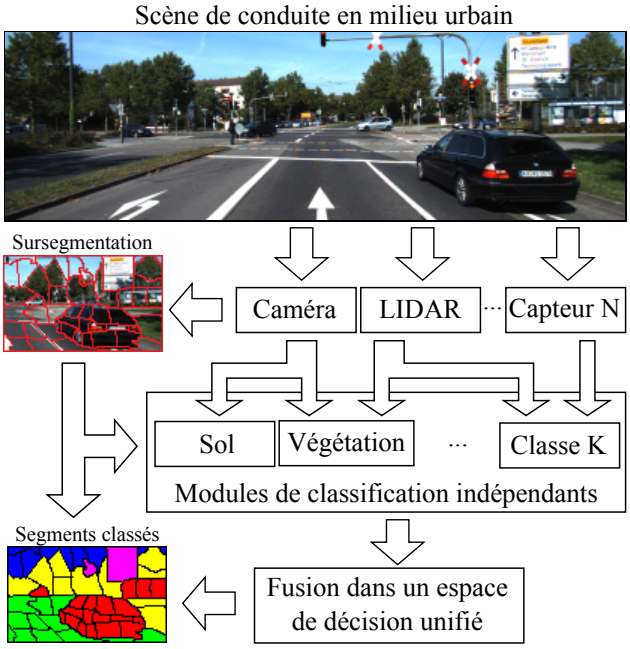


FIGURE 1 – Vue d'ensemble du schéma de fusion. N capteurs, dont une caméra, observent la scène et fournissent en données K modules indépendants. Les résultats de classification sont ensuite fusionnés dans un espace de décision unique au niveau d'une image sursegmentée.

Keywords

Information fusion, driving scenes understanding, theory of belief functions, intelligent vehicles.

1 Introduction

La compréhension de scènes routières est une tâche complexe qui amène à considérer de nombreuses sous-tâches, allant de la détection d'objets à la localisation en passant par la reconstruction 3D. Tous ces problèmes ont fait l'objet de nombreux travaux de recherche durant les dernières décennies. Malheureusement, chacune de ces tâches est

souvent traitée de manière indépendante et isolée, en n'utilisant qu'un ou plusieurs capteurs spécifiques. Afin de pouvoir profiter au mieux de tous les travaux existants, il devient essentiel de pouvoir fusionner proprement toutes les sources d'information à disposition.

Plusieurs questions importantes se posent alors. Comment un détecteur de végétation peut-il, par exemple, aider un détecteur de piéton et vice versa ? Comment les données issues d'un capteur LIDAR (capteur laser), qui ne perçoit qu'un ensemble discret d'impacts réfléchis par des obstacles, peuvent-elles être fusionnées avec un module de détection de ciel basé sur une caméra ? Comment inclure de nouveaux capteurs ou de nouvelles classes d'objets ?

Plus généralement, deux buts critiques doivent être atteints. Le premier est de pouvoir combiner plusieurs modules traitant des classes d'objets différentes et d'être assez flexible pour inclure de nouvelles classes. Le deuxième but est de pouvoir représenter, dans un espace commun, les données issues de capteurs pouvant observer l'environnement de manières très différentes.

1.1 Travaux antérieurs

Dans le domaine des véhicules intelligents, les caméras et les LIDARs sont les capteurs les plus souvent utilisés pour la perception. Les capteurs LIDAR ont, par exemple, été utilisés pour détecter les structures statiques et les objets en mouvement, notamment dans des contextes de construction de cartes [1, 2] ou de grilles d'occupation [3]. Les caméras ont, quant à elles, été considérées dans un champ d'applications beaucoup plus large. La détection de piétons est l'un des cas les plus étudiés [4]. Des travaux, plus généraux, de classifications multi-classes pour la compréhension de scènes urbaines ont également été menés [5]. Ces derniers utilisent parfois un système stéréoscopique permettant d'avoir une information de profondeur [6] et peuvent notamment servir à détecter les obstacles mais aussi les zones navigables [7].

En ce qui concerne l'aspect fusion, de nombreuses méthodes basées sur des systèmes multi-capteurs utilisent une approche de type «régions d'intérêt». Le principe est d'utiliser un premier capteur, par exemple un LIDAR [8], pour sélectionner un ensemble de régions candidates, qui seront ensuite analysées plus finement à l'aide d'autres sources d'information. D'autres approches passent par l'estimation de la configuration géométrique de la scène, en calculant, par exemple, le plan du sol ou le point de fuite [9, 10], afin de contraindre la recherche d'objets. Enfin, un autre type de fusion consiste à combiner différentes caractéristiques visuelles [4] et/ou géométriques [5, 6] afin d'avoir un plus grand pouvoir discriminant. Ces méthodes de fusion sont souvent construites autour d'un problème spécifique prédéfini. Il devient alors difficile d'inclure de nouvelles classes d'objets ou caractéristiques sans avoir à réentraîner tout le système. De même, la fusion avec un nouveau capteur ou module de traitement n'est pas envisageable.

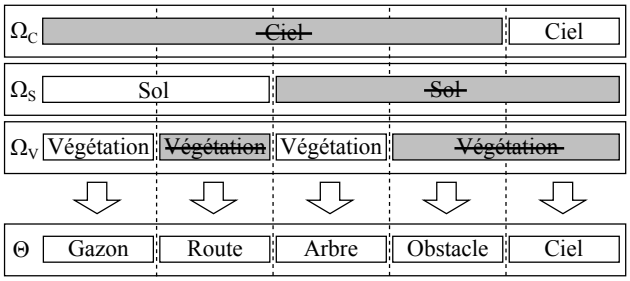


FIGURE 2 – Illustration d'une fusion multi-classes. Les trois premiers blocs représentent trois différentes décompositions du monde. Les blocs gris correspondent aux complémentaires des blocs blancs. En les intersectant, on obtient un raffinement commun. La classe «obstacle» représente, en fait, tout ce qui n'est ni ciel, ni sol, ni végétation.

1.2 Contributions

Dans cet article, nous nous attachons à construire un système permettant de combiner différents modules de traitement sans contrainte sur leur tâche spécifique ni les capteurs dont ils dépendent. Cette flexibilité permet non seulement de pouvoir envisager de nouvelles classes à venir mais aussi de restreindre les classes à analyser. Ainsi, il ne sera pas nécessaire de définir à l'avance une liste exhaustive des objets pouvant apparaître dans la scène, ce qui impliquerait de devoir construire un détecteur pour chacun d'eux. Pour ce faire, la théorie des fonctions de croyance ou théorie de Dempster-Shafer [11] est utilisée.
Nous montrons également comment combiner des sources d'information dont les représentations peuvent être de natures différentes. Dans ce but, nous formulons le problème comme celui de l'annotation d'image en utilisant une image sursegmentée.
Une vue d'ensemble du système est illustrée sur la figure 1, plusieurs capteurs, dont une caméra, observent la scène et fournissent en données différents modules de traitement indépendants. Ces derniers peuvent indépendamment utiliser les données provenant d'un ou plusieurs capteurs. Les résultats de classification de ces modules, qui concernent a priori des classes d'objets différentes, sont tout d'abord projetés dans un espace de décision commun avant d'être fusionnés au niveau d'une image sursegmentée.
Nous montrerons comment ce système peut être utilisé dans la pratique, en considérant un système comprenant une caméra stéréo et un capteur LIDAR. Plusieurs modules seront décrits, tout d'abord pour la détection du sol, puis pour un problème plus large incluant la végétation et le ciel. Une validation expérimentale est menée sur des données réelles provenant de la base de données KITTI Vision Benchmark Suite [12].

2 Théorie des fonctions de croyance

Pour montrer l'utilité de la théorie des fonctions de croyance, prenons l'exemple de la fusion entre un détec-

teur de sol, un détecteur de ciel et un détecteur de végétation. Comme l'illustre la figure 2, en intersectant ces trois classes initiales, on peut obtenir de nouvelles sous-classes. La classe «sol» peut, par exemple, être divisée en «gazon» et «route» en l'intersectant avec la classe «végétation».
Une connaissance spécifique à la classe «sol» ne donne, a priori, aucune information sur les classes «gazon» et «route», si ce n'est qu'elles sont toutes les deux aussi plausibles l'une que l'autre. Il est notamment injustifié de distribuer uniformément la connaissance sur la classe «sol» aux classes «gazon» et «route», car une connaissance artificielle quant aux deux nouvelles est créée. Il faut donc pouvoir raisonner au niveau d'ensembles de classes, ce que permet la théorie des fonctions de croyance.

Quelques définitions Soit Ω un ensemble de classes mutuellement exclusives, appelé *cadre de discernement*, représentant l'ensemble de toutes les classes d'objets possibles. On appelle *fonction de masse* [11] sur Ω, une fonction $m : 2^\Omega \to [0, 1]$ telle que :

$$m(\emptyset) = 0, \qquad \sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Pour une variable y à valeur dans Ω, la croyance relative à son appartenance aux différentes parties de Ω peut être modélisée par une fonction de masse m. Étant donné un sous-ensemble A de Ω, la quantité $m(A)$, qu'on appellera la masse sur A, peut être interprétée comme la croyance allouée spécifiquement à l'hypothèse $y \in A$.
Tout sous-ensemble $A \subseteq \Omega$, tel que $m(A) > 0$, est appelé *élément focal* de m. On dira que m est une fonction de masse *catégorique* sur A, si A est l'unique élément focal de m. En particulier, si $A = \Omega$, m représente l'ignorance totale, l'hypothèse $y \in \Omega$ étant supposée toujours vraie. On peut remarquer qu'une fonction de masse n'ayant que des singletons comme éléments focaux représente exactement une distribution de probabilité. La théorie des fonctions de croyance est donc une généralisation des probabilités bayesiennes classiques.

Raffinement et grossissement Partant, d'un cadre de discernement Ω, on peut définir un raffinement Θ en partitionnant un ou plusieurs éléments de Ω. Sur l'exemple de la figure 2, le cadre de discernement Θ est un raffinement commun à Ω_C, Ω_S et Ω_V. Le raffinement de Ω en Θ est défini par une application $\rho : 2^\Omega \to 2^\Theta$ telle que :

1. $\{\rho(\{\omega\}),\ \omega \in \Omega\} \subseteq 2^\Theta$ est une partition de Θ;
2. $\forall A \in \Omega,\ \rho(A) = \bigcup_{\omega \in A} \rho(\{\omega\})$.

Par exemple, pour raffiner $\Omega_S = \{Sol, \overline{Sol}\}$ en $\Theta = \{Gazon, Route, Arbre, Obstacle, Ciel\}$, on définit $\rho : 2^\Omega \to 2^\Theta$ par :

$$\begin{aligned} \rho(\{Sol\}) &= \{Gazon, Route\}, \\ \rho(\{\overline{Sol}\}) &= \{Arbre, Obstacle, Ciel\}. \end{aligned}$$

La deuxième propriété sur ρ imposera alors naturellement $\rho(\Omega) = \rho(\Theta)$. On peut alors transformer une fonction de masse m^Ω définie sur Ω en une fonction de masse m^Θ définie sur Θ en posant pour tout $B \in \Theta$:

$$m^\Theta(B) = \begin{cases} m^\Omega(A) & \text{si } \exists A \in \Omega,\ B = \rho(A), \\ 0 & \text{sinon.} \end{cases} \tag{2}$$

Une masse initialement assignée à $\{Sol\}$ sera, par exemple, transférée sur $\{Gazon, Route\}$ au niveau de Θ. Réciproquement, l'opération inverse, appelée grossissement, peut également être définie de manière similaire.

Combinaison de fonctions de croyance Étant donné un cadre de discernement Ω et deux fonctions de masse m_1, m_2, construites à partir de sources d'information indépendantes, elles peuvent être combinées pour former une nouvelle masse $m_{1,2} = m_1 \oplus m_2$, en utilisant la règle de combinaison de Dempster :

$$\begin{aligned} m_{1,2}(\emptyset) &= 0, \\ m_{1,2}(A) &= \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \end{aligned} \tag{3}$$

avec $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$, qui est une mesure du conflit entre les deux sources d'information. Cette règle de combinaison est associative et commutative, l'ordre dans lequel les sources sont combinées n'a donc pas d'influence sur le résultat final. Pour combiner deux fonctions de masse définies sur des cadres de discernement différents, il suffit de trouver un raffinement commun puis d'utiliser la règle de Dempster. Si nécessaire, un grossissement peut être utilisé pour se ramener à un des deux cadres de discernement initiaux.

Affaiblissement Il est parfois intéressant de pouvoir affaiblir une fonction de masse, notamment lorsqu'on dispose d'une mesure de sa fiabilité. L'affaiblissement d'une fonction de masse m par un facteur $\alpha \in [0, 1]$ s'exprime comme :

$$\begin{aligned} {}^\alpha m(A) &= (1-\alpha) m(A), \quad \forall A \subset \Omega, \\ {}^\alpha m(\Omega) &= (1-\alpha) m(\Omega) + \alpha. \end{aligned} \tag{4}$$

Autrement dit, la masse sur tous les éléments focaux est diminuée d'un facteur $1 - \alpha$ et le reste est transféré sur l'ignorance.

Prise de décision Enfin, pour passer d'une fonction de masse à une prise de décision, plusieurs approches sont possibles. La plus répandue consiste à calculer la probabilité pignistique [13] $BetP$ pour tous les éléments $\omega \in \Omega$:

$$BetP(\omega) = \sum_{A \subseteq \Omega,\ \omega \in A} \frac{m(A)}{|A|}. \tag{5}$$

Puis, le singleton avec la plus grande probabilité pignistique est sélectionné. Cependant, cette approche n'est pas du tout adaptée dans un contexte où le cadre de discernement peut être changé, en particulier raffiné. Le calcul de

la probabilité pignistique utilisant les cardinaux des parties de Ω, il est directement influencé par le découpage du cadre de discernement.
Une autre approche consiste à choisir le singleton avec la plus grande mesure de plausibilité pl, qui est définie pour tout $A \subseteq \Omega$ par :

$$pl(A) = \sum_{B \subseteq \Omega,\ B \cap A \neq \emptyset} m(B). \quad (6)$$

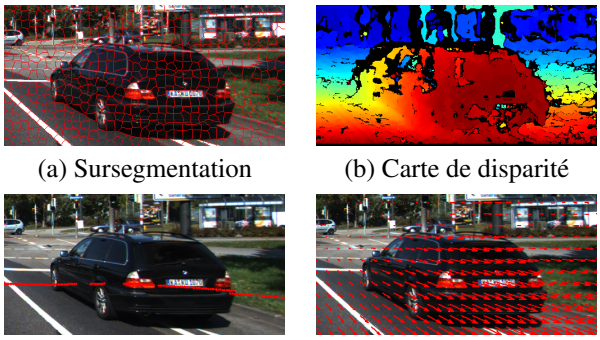
Cette approche est beaucoup plus adaptée car la plausibilité accordée à un sous-ensemble donné reste inchangée même si le cadre de discernement est modifié. Elle sera donc utilisée dans nos travaux.

3 Annotation d'images sursegmentées

Comme expliqué en introduction, non seulement il est nécessaire de raisonner avec des classes différentes mais également avec des représentations différentes des données. Étant dans un contexte d'aide à la conduite où le but est d'avertir le conducteur de dangers potentiels, il est raisonnable de travailler au niveau d'une image. L'image acquise par une caméra reflétant, en effet, ce que perçoit le conducteur.
Raisonner au niveau du pixel est souvent trop local et difficile, tandis que raisonner au niveau des objets (e.g., en utilisant des boîtes englobantes) est inadapté pour certaines classes d'objets comme la route. Nous avons choisi un niveau intermédiaire en sursegmentant l'image. De nombreux algorithmes de sursegmentation peuvent être trouvés dans la littérature [14, 15]. L'algorithme TurboPixels de Levinshtein et al. [16] a été choisi car il propose une sursegmentation en forme de grille régulière. La formulation ne dépend, cependant, pas de l'algorithme de sursegmentation utilisé. Il peut même être intéressant de combiner plusieurs segmentations comme l'ont fait Mathevet et al. [17] en utilisant la théorie des fonctions de croyance.
La tâche commune à tous les modules devient alors de classer chaque segment de l'image. Peu importe la représentation en amont, le résultat doit être projeté au niveau de l'image. La théorie des fonctions de croyance permettant de représenter l'ignorance, il n'est pas nécessaire, pour chaque module, de classer tous les segments de l'image.

4 Application à la compréhension de scènes routières

Nous appliquons notre schéma de fusion à un système composé d'une caméra stéréo et d'un LIDAR, que nous supposons calibré. Plusieurs modules indépendants traitent les données issues de ces capteurs afin de classer les segments de l'image. Les informations 3D capturées par la caméra stéréo et le LIDAR sont utilisées dans un premier temps pour détecter le sol. Elles seront également utilisées, par la suite, pour la détection de tout ce qui n'est pas le ciel. Un module basé sur l'analyse de texture est également



(a) Sursegmentation (b) Carte de disparité

(c) Points laser (d) Flot optique

FIGURE 3 – Données d'entrée du système de fusion multi-capteurs.

envisagé pour la détection du ciel , de la route et de la végétation. Enfin, une propagation temporelle est utilisée afin de lier deux images consécutives. Les données d'entrée des différents modules sont représentées sur la figure 3.

4.1 Fonction de masse à partir d'un modèle

Un problème de classification peut souvent être vu comme la recherche d'une correspondance entre un modèle préappris M d'une classe C et une observation X d'un objet S (qui dans notre cas est un segment d'image). À partir d'une mesure $d(X, M)$ entre une observation et un modèle, la classe de l'objet S peut être inférée.
Pour construire une fonction de masse m sur un cadre de discernement $\Omega = \{C, \overline{C}\}$, où $\overline{C}$ comprend tout ce qui n'appartient pas à C, une approche générale consiste à définir deux fonctions de masse. Une première, m^-, qui va attribuer de la croyance à la classe C si $d(X, M)$ est petite et une seconde, m^+, qui va, au contraire, attribuer de la croyance à $\overline{C}$ si $d(X, M)$ est grande.
Il est important de noter que, dans certains cas, on ne peut inférer la classe de S que lorsque la valeur de $d(X, M)$ est grande, tandis que rien ne peut être dit dans le cas contraire ou inversement. Dans ce genre de situation, une seule des deux fonctions masses m^- et m^+ est utilisée, sinon elles sont combinées par la règle de Dempster.
Des formes générales pour m^- et m^+, inspirées de celle proposée par Denœux [18], sont :

$$\begin{aligned} m^-(\{C\}) &= e^{-\gamma^- \left(\frac{d}{d^- - d}\right)^\beta} \text{ si } d < d^-,\ 0 \text{ sinon}, \\ m^-(\Omega) &= 1 - m^-(\{C\}), \end{aligned} \quad (7)$$

et

$$\begin{aligned} m^+(\{\overline{C}\}) &= e^{-\gamma^+ \left(\frac{d^+}{d - d^+}\right)^\beta} \text{ si } d > d^+,\ 0 \text{ sinon}, \\ m^+(\Omega) &= 1 - m^+(\{\overline{C}\}). \end{aligned} \quad (8)$$

Les seuils d^- et d^+ définissent les valeurs au dessous et au dessus desquelles de la masse peut être attribuée à $\{C\}$ et $\{\overline{C}\}$. Le paramètre $\beta \in \{1, 2, \ldots\}$, qui peut être arbitrairement fixé à 1 ou 2, comme suggéré dans [18], et $\gamma > 0$

reflètent l'influence de la distance quant à la masse allouée. La fonction de masse combinée $m = m^- \oplus m^+$ peut, finalement, être affaiblie par un facteur α si nécessaire.
À partir d'une base d'entraînement $\{(X_i, c_i)\}_{1 \leq i \leq n}$, où $c_i \in \{C, \overline{C}\}$ est la classe de l'observation X_i, les paramètres peuvent être choisis pour minimiser la fonction de perte suivante :

$$L = \sum_{i=1}^{n} 1 - pl_i(\{c_i\}) + pl_i(\{\overline{c_i}\}), \quad (9)$$

où pl_i est la plausibilité (6) associée à l'observation X_i. La perte $L_i = 1 - pl_i(\{c_i\}) + pl_i(\{\overline{c_i}\})$ possède les propriétées suivantes :

$$\begin{aligned} m_i(\{c_i\}) \to 1 &\Rightarrow L_i \to 0, \\ m_i(\{\overline{c_i}\}) \to 1 &\Rightarrow L_i \to 2, \\ m_i(\Omega) \to 1 &\Rightarrow L_i \to 1. \end{aligned}$$

La fonction de perte attribue un coût élevé pour de la masse incorrectement allouée et un coût intermédiaire pour l'ignorance.

4.2 Classification basée stéréo

Une caméra stéréo permet d'estimer la profondeur de chaque pixel d'une image. L'information 3D est générée par le calcul d'une carte de disparité (Fig.3(b)), qui peut éventuellement être incomplète et erronée. Dans notre étude, nous avons utilisé l'approche semi-globale proposée par Hirschmüller [19].
Dans une premier temps, les informations 3D sont utilisées pour détecter le sol. Pour ce faire, nous utilisons un estimateur robuste (*RANSAC*) pour détecter le plan du sol, sous l'hypothèse que le sol soit bien plan. Des modèles plus complexes existent également dans la littérature pour traiter les cas de sol non plan. À partir de là, un segment d'image est classé comme sol ou non-sol selon sa distance par rapport au plan estimé.
Le cadre de discernement est donc $\Omega = \{Sol, \overline{Sol}\}$. Le modèle de la classe «sol» est tout simplement l'équation du plan du sol, qu'on notera Π. Un segment d'image S est représenté par un ensemble de n points $X = \{p_1, p_2, \ldots, p_k, p^*_{k+1}, \ldots, p^*_n\}$, où les points p^*_i sont ceux dont la disparité n'a pu être calculée. La distance entre l'observation X et le modèle Π est alors définie comme la distance moyenne des points valides p_i au plan Π :

$$d(X, \Pi) = \frac{1}{k} \sum_{i=1}^{k} d(p_i, \Pi), \quad (10)$$

où $d(p_i, \Pi)$ est la distance euclidienne entre le point p_i et Π. Cette distance est alors utilisée pour construire une fonction de masse en utilisant les formules (7) et (8). Le segment S pouvant n'être que partiellement visible, la fonction de masse est affaiblie par un coefficient $\alpha = 1 - k/n$, représentant la proportion de points dont la disparité n'a pu être estimée par rapport au nombre total de points dans S. Ainsi, si la disparité d'aucun point de S n'a pu être estimée, on se retrouve avec une masse vide $m(\Omega) = 1$.
Des informations supplémentaires peuvent être tirées de l'équation du plan du sol, comme par exemple la ligne d'horizon. Cette dernière aurait également pu être estimée à partir d'une seule image en passant par des méthodes d'estimation de point de fuite. La connaissance de la ligne d'horizon permet alors de dire que tout segment au dessus de celle-ci appartient forcément à la classe «non-sol». Ce qui est représenté par une masse catégorique $m(\{\overline{Sol}\}) = 1$, qui peut être combinée avec la fonction de masse précédemment calculée en utilisant la règle de combinaison de Dempster (3).
L'analyse peut également être étendue à d'autres cadres de discernement, par exemple avec $\Theta = \{Ciel, \overline{Ciel}\}$. En effet, on sait que tout segment en dessous de la ligne d'horizon ne peut appartenir à la classe «ciel». De plus, les points du ciel pouvant être supposés à des distances presque infinies, i.e. avec une disparité nulle, les segments de disparité moyenne strictement positive, ou plus grande qu'un certain seuil, peuvent également être classées comme «non-ciel». Pour combiner cette nouvelle source d'information, on définit un nouveau cadre de discernement, commun à Ω et Θ, $\Psi = \{Sol, Ciel, \overline{Sol \cup Ciel}\}$.

4.3 Classification basée LIDAR

Un capteur LIDAR fournit, comme une caméra stéréo, des informations 3D. La figure 3(c) montre les impacts lasers projetés sur l'image. Ainsi, comme dans le cas d'une caméra stéréo, un segment S est perçu comme un ensemble de k points 3D, $X = \{p_1, \ldots, p_k\}$, pouvant éventuellement être vide. Pour les segments comportant des impacts lasers, on utilise la même approche que dans le cas stéréo, en reprenant la distance (10) et la formule (**??**).
De plus, tout l'espace entre un impact laser et l'origine du capteur LIDAR est vide et est apparenté au sol. Ainsi, pour tous les segments S par lesquels sont passés des rayons laser, on associe la fonction de masse suivante :

$$\begin{aligned} m(\{Sol\}) &= k/n \\ m(\{\overline{Sol}\}) &= 0 \\ m(\Omega) &= 1 - k/n \end{aligned} \quad , \quad (11)$$

où k est le nombre de rayons mesurés passant à travers S et n le nombre maximal de rayons ayant potentiellement pu traverser S. On peut remarquer que cette fonction de masse est, en fait, l'affaiblissement d'une masse catégorique sur $\{Sol\}$ par un coefficient $\alpha = 1 - k/n$.

4.4 Classification basée sur la texture

La texture est une information importante pour les tâches de classification. Pour encoder l'information de texture, la transformation de Walsh-Hadamard est utilisée, en suivant l'approche de Wojek et Schiele [20]. Pour construire un modèle, pour une classe donnée, une approche *sac de mots* [21] est employée. Les caractéristiques de textures

sont tout d'abord quantifiées en un ensemble de N *textons* ; un modèle est alors simplement un histogramme normalisé $H = (h_1, \ldots, h_N)$, où m_i est la fréquence d'apparition du $i^{\text{ème}}$ texton pour la classe en question. Chaque segment est également observé sous la forme d'un histogramme $X = (x_1, \ldots, x_N)$, sa distance au modèle H peut alors être calculée à l'aide d'une distance entre histogrammes, comme la distance χ^2 par exemple :

$$d(X, H) \quad = \quad \frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - h_i)^2}{x_i + h_i}. \qquad (12)$$

Cette approche permet de pouvoir construire un modèle pour chaque classe indépendamment les unes des autres. Seuls des exemples positifs sont nécessaires pour apprendre le modèle. On peut alors se dispenser de considérer toutes les classes d'objets possibles, d'autant plus que pour certains types d'objets, l'information de texture n'est pas assez discriminante. Dans nos travaux, nous nous limitons aux classes «végétation», «ciel» et «route», ces trois classes étant bien adaptées à une analyse de texture.
En travaillant avec des informations de texture, il peut arriver qu'une petite distance entre une observation et un modèle ne soit pas suffisant pour inférer la classe de l'objet en question. Par exemple, la façade blanche d'un bâtiment est très proche, en terme d'apparence, du ciel, qui apparaît souvent blanc sur l'image. Toutefois, lorsque la texture est éloignée de celle du ciel, on peut déduire que l'objet n'appartient pas à la classe «ciel». Ainsi, il est plus prudent, ici, de ne considérer que les masses m^+.

4.5 Propagation temporelle

Enfin, un dernier module de traitement, propage le résultat de classification d'un instant t à un instant $t+1$ à l'aide du flot optique (Fig. 3(d)). Ce dernier est calculé à partir de deux images consécutives en utilisant les travaux de Werlberget et al. [22]. À chaque segment S_t de l'image à l'instant t est associé un segment S_{t+1} à l'instant $t+1$, défini comme celui pointé par le flot optique médian des pixels de S_t. La fonction de masse associée à S_t est alors directement transférée à S_{t+1} :

$$\forall A \subseteq \Omega, \qquad m_{t+1}(A) = m_t(A). \qquad (13)$$

Elle est ensuite affaiblie par un facteur α correspondant au ratio de pixels de S_t ne pointant pas vers S_{t+1}. Enfin, si plusieurs segments pointent vers le même segment à $t+1$, les fonctions de masse associées sont simplement combinées par la règle de Dempster.

5 Résultats expérimentaux

La base de données KITTI [12] a été utilisée pour valider l'approche proposée. La caméra stéréo couleur ainsi que le LIDAR Velodyne ont été utilisés comme capteurs. Toutefois, une seul nappe du LIDAR Velodyne a été considérée afin de simuler un capteur LIDAR simple nappe, communément utilisé en robotique mobile.

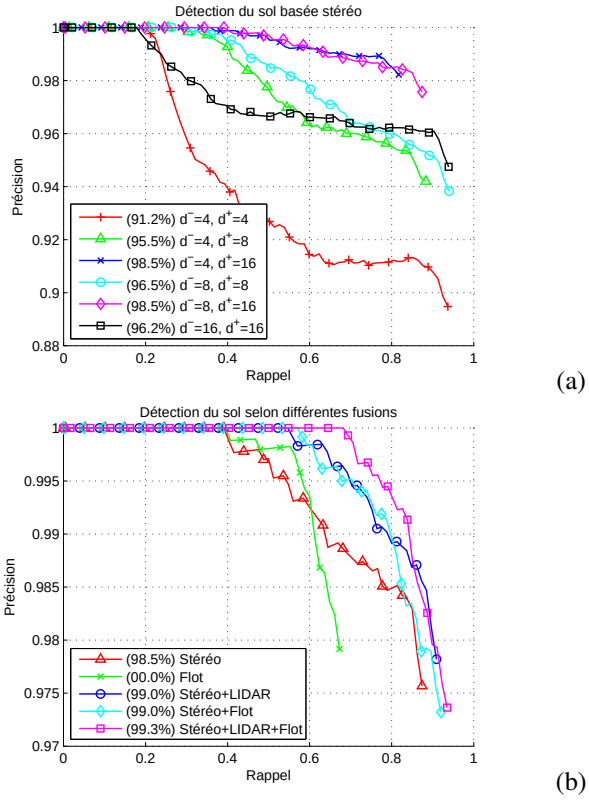


FIGURE 4 – Courbes précision/rappel pour la détection du sol. Les valeurs entre parenthèses correspondent à la précision pour un taux de rappel de 80%. (a) Performances du module basé stéréo pour différentes valeurs de d^- et d^+. (b) Performances en combinant différents modules.

Classification basée sur la texture

	Gazon	Route	Arbre	Obst	Ciel	Rappel
Gazon						0
Route		66		33.8	0.2	50.2
Arbre						0
Obst		14.3		84.1	1.6	49.3
Ciel		0		18.4	81.6	80.5

(a)

Classification après fusion

	Gazon	Route	Arbre	Obst	Ciel	Rappel
Gazon	81.1	3.8	15.1	0	0	40.6
Route	7.9	89.1	0.3	2.7	0	78.8
Arbre	2.5	0	94.4	3.1	0	86.7
Obst	0.8	2.3	9.1	86.8	1	52.6
Ciel	0	0	0	18.4	81.6	80.5

(b)

FIGURE 5 – Matrices de confusion pour la classification multi-classe. (a) Résultats de la classification basée uniquement sur l'analyse de texture. (b) Résultats après fusion avec les modules stéréo, LIDAR et flot optique.

Les paramètres pour chaque module ont été choisis en testant différentes configurations. La figure 4(a) montre l'influence des seuils d^- et d^+ sur la détection du sol basée sur le module stéréo. Les valeurs $d^- = 8$ et $d^+ = 16$ semblent être un bon compromis entre précision et rappel. Pour un taux de rappel de 80%, nous obtenons une précision de $98,5\%$. Nous remarquons que le taux de rappel

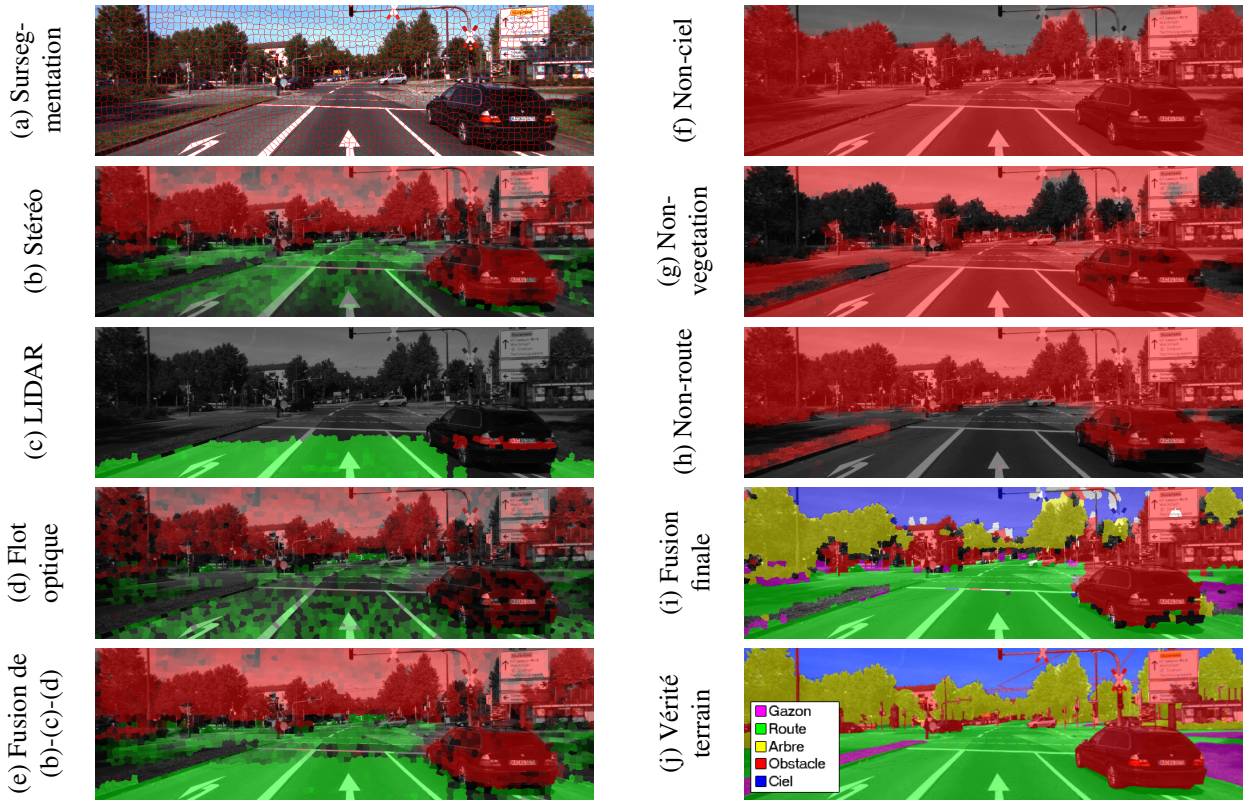


FIGURE 6 – (a) Image sursegmentée. (b-e) Résultats de détection du sol avec $\Omega = \{Sol, \overline{Sol}\}$. L'intensité en vert représente la masse affectée à $\{Sol\}$ tandis que le rouge représente la masse sur $\{\overline{Sol}\}$. En l'absence de couleur, la masse est assignée à Ω. (f-h) Analyse par texture des classes ciel, vegetation et route, les masses sont uniquement allouées aux classes complémentaires (représentées en rouge). (i) Fusion de tous les modules, le code couleur est celui de la vérité terrain (j).

maximal est atteint pour les configurations où $d^- = d^+$. En effet, lorsque ce n'est pas le cas, il y a une zone d'ignorance dans laquelle aucune décision ne peut être prise. En permettant de ne pas prendre de décision dans les cas ambiguës, la précision est grandement améliorée. Quant aux paramètres β, γ^- et γ^+, ils n'ont pas d'influence sur la précision ni le rappel lorsque le module est considéré tout seul. β est arbitrairement fixé à 2 tandis que γ^- et γ^+ sont choisis par rapport à la fonction de perte (9).

En combinant plusieurs détecteurs de sol, issus de la stéréo, du LIDAR et du flot optique, les performances sont clairement améliorées comme illustré dans la figure 4(b). Nous remarquons que la combinaison de plusieurs sources est systématiquement meilleure que chacune des sources prise individuellement. Les figures 6(b-e) illustrent les différentes masses affectées aux classes «sol» et «non-sol».

Le module basé sur l'analyse de texture a été utilisé pour la détection de la route, de la végétation et du ciel. La figure 5(a) montre la matrice de confusion de la classification multi-classe et les figures 6(f-h) illustrent les résultats obtenus. Utilisé seul, ce module ne peut différencier le gazon des arbres. Par contre, cela devient possible en le combinant avec les détecteurs de sol. La figure 5(b) montre la matrice de confusion du système complet. Nous voyons encore une fois une amélioration des résultats.

Nous rappelons que dans certain cas, aucune décision ne peut être prise (cas de l'ignorance), ce qui explique que le taux de rappel soit inférieur à la diagonale de la matrice de confusion. Nous pouvons, par exemple, remarquer que le taux de rappel de la classe «gazon» est particulièrement bas. Cela s'explique par le fait que le gazon est parfois légèrement surélevé par rapport au sol. Le gazon n'est alors même pas classé comme sol, comme nous pouvons le voir sur la partie en bas à gauche de l'image 6(j) où les régions non colorées sont celles n'ayant pas été classées.

6 Conclusions et perspectives

Nous avons proposé un schéma original de fusion d'informations basé sur une sursegmentation et la théorie des fonctions de croyance. Il est suffisamment flexible pour pouvoir rajouter de nouvelles classes d'objets, de nouveaux modules de traitement et de nouveaux capteurs.

Des travaux futurs seront menés afin d'ajouter de nouvelles classes comme les piétons ou les véhicules, en adaptant notamment des méthodes basées sur l'utilisation de fenêtres glissantes. En combinant avec une information de profondeur, l'information au niveau d'une boîte englobante pourra, par exemple, être ramenée à celui des segments de l'image. De nouvelles sources d'information comme

le GPS ou les cartes seront également considérées pour la détection d'objets en mouvement. Enfin, une approche globale sera également étudiée afin de fusionner des segments voisins appartenant au même objet, ce qui permettra d'avoir une compréhension à plus haut niveau de la scène.

7 Remerciements

Ce travail est soutenu et financé par le programme Cai Yuanpei, accordé par le Ministère chinois de l'Éducation (MOE) et les Ministères français des Affaires Étrangères et Européennes (MAEE) et de l'Enseignement Supérieur et de la Recherche (MESR), ainsi que par le projet Blanc International ANR-NSFC franco-chinois PRETIV.

Références

[1] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics (Intelligent Robotics and Autonomous Agents)*. Cambridge, Massachusetts : The MIT Press, 2005.

[2] C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *International Journal of Robotics Research*, vol. 26, no. 1, pp. 889–916, 2007.

[3] J. Moras, V. Cherfaoui, and P. Bonnifait, "Moving objects detection by conflict analysis in evidential grids," in *IEEE Intelligent Vehicles Symposium*, (Baden-Baden, Germany), pp. 1120–1125, 2011.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection : an evaluation of the state of the art," *PAMI*, vol. 34, no. 4, pp. 743–761, 2011.

[5] A. Ess, T. Müller, H. Grabner, and L. V. Gool, "Segmentation based urban traffic scene understanding," in *BMVC*, (London, UK), pp. 1–11, 2009.

[6] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr, "Joint optimization for object class segmentation and dense stereo reconstruction," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 122–133, 2012.

[7] H. Badino, U. Franke, and R. Mester, "Free space computation using stochastic occupancy grids and dynamic programming," in *ICCV Workshop on Dynamical Vision*, (Rio de Janeiro, Brazil), 2007.

[8] S. Rodríguez, V. Frémont, P. Bonnifait, and V. Cherfaoui, "Multi-modal object detection and localization for high integrity driving assistance," *Machine Vision and Applications*, vol. 14, pp. 1–16, 2011.

[9] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool, "Dynamic 3d scene analysis from a moving vehicle," in *CVPR*, (Minneapolis, USA), pp. 1–8, 2007.

[10] A. Geiger, C. Wojek, and R. Urtasun, "Joint 3d estimation of objects and scene layout," in *NIPS*, (Granada, Spain), pp. 1467–1475, 2011.

[11] G. Shafer, *A mathematical theory of evidence*. Princeton, New Jersey : Princeton University Press, 1976.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving ? The KITTI vision benchmark suite," in *CVPR*, (Providence, USA), pp. 3354–3361, 2012.

[13] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–243, 1994.

[14] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[15] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.

[16] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi, "TurboPixels : Fast superpixels using geometric flows," *PAMI*, vol. 31, no. 12, pp. 2290–2297, 2009.

[17] S. Mathevet, L. Trassoudaine, P. Checchin, and J. Alizon, "Combinaison de segmentations en régions," *Traitement du Signal*, vol. 16, no. 2, pp. 93–104, 1999.

[18] T. Denœux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[19] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *PAMI*, vol. 30, no. 2, pp. 328–341, 2008.

[20] C. Wojek and B. Schiele, "A dynamic conditional random field model for joint labeling of object and scene classes," in *ECCV*, (Marseille, France), pp. 733–747, 2008.

[21] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories : a comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.

[22] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *CVPR*, (San Francisco, USA), pp. 2464–2471, 2010.