Méthode et système d'assistance à la navigation de personne basés sur la perception sonore d'une scène visuelle

Florian Scalvini

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE A l'UNIVERSITÉ DE BOURGOGNE

École doctorale n^{o} 37

Sciences Physiques pour l'Ingénieur et Microtechniques

Doctorat d'Informatique et Instrumentation de l'Image

Par

M. SCALVINI Florian

Méthode et système d'assistance à la navigation de personne basés sur la perception sonore d'une scène visuelle.

Thèse présentée et soutenue à Dijon, le 5 avril 2024

Composition du Jury :

Pr. Christophe DUCOTTET	Professeur, Université Jean Monnet	Président / Examinateur
Pr. François BERRY	Professeur, Université Clermont Auvergne	Rapporteur
Pr. Denis PELLERIN	Professeur, Université Grenoble Alpes	Rapporteur
Pr. Julien DUBOIS	Professeur, Université de Bourgogne	Directeur de thèse
Dr. Maxime AMBARD	Maître de Conférences, Université de Bourgogne	Co-encadrant de thèse
Dr. Cyrille MIGNIOT	Maître de Conférences, Université de Bourgogne	Co-encadrant de thèse

Remerciements

La rédaction de ce manuscrit de thèse marque la fin d'un chapitre intense et enrichissant de trois ans de doctorat. Cette période, bien que parfois compliquée et stressante, se révèle rétrospectivement très agréable. La plénitude ressentie en écrivant ces lignes témoigne du chemin parcouru aussi bien professionnellement que personnellement. Ces années auront été marquées par le soutien et l'accompagnement de plusieurs personnes que je souhaiterais remercier.

Je souhaite exprimer ma profonde reconnaissance à Monsieur François Berry et Monsieur Denis Pellerin, respectivement Professeur à l'université de Clermont-Auvergne et à l'université de Grenoble-Alpes, pour m'avoir fait l'honneur d'être rapporteurs de ma thèse. De même, je souhaite remercier chaleureusement Monsieur Christophe Ducottet, Professeur à l'Université Jean Monnet, pour avoir accepté d'être examinateur lors de la soutenance de thèse.

Je voudrais remercier la région Bourgogne Franche-Comté pour avoir accompagné pendant ces trois dernières années le projet 3DGS avec un financement Envergure (BG0027904).

Je voudrais exprimer individuellement ma profonde gratitude envers l'ensemble de mes encadrants de thèse pour leur formidable accompagnement, en commençant par Monsieur Julien Dubois, directeur de thèse. Vous m'avez fourni un soutien inestimable, depuis mes premières années à l'école d'ingénieur jusqu'à aujourd'hui. Votre confiance en moi, durant l'ensemble de ces années, m'ont permis de surmonter mes doutes.

Je remercie chaleureusement Monsieur Cyrille Migniot pour sa constante disponibilité. La bienveillance quotidienne dont vous avez fait preuve ainsi que vos conseils précieux, particulièrement lors de la rédaction d'articles scientifiques, m'ont été extrêmement bénéfiques et m'accompagneront longtemps.

Je voudrais exprimer également ma gratitude envers Monsieur Maxime Ambard sans qui cette thèse captivante n'aurait pas pu avoir lieu. Le sujet innovant et passionnant a grandement contribué à mon épanouissement. Ce doctorat à la croisée de la vision par ordinateur et de la psychologie humaine, en particulier sur les problématiques liées au handicap, a été pour moi une aventure exceptionnellement enrichissante, tant sur le plan professionnel que personnel.

Je ne saurais poursuivre ces remerciements sans évoquer les membres permanents, les postdoctorants, les doctorants et les stagiaires présents au sein du laboratoire ImViA pour leurs accueils chaleureux et leurs bienveillances. Un merci tout particulier à Sean, Joffrey et Hermes pour les innombrables parties de SSBU qui ont animé nos

pauses-déjeuner, mais également à Mathilde pour m'avoir assistée pour la conception des prototypes expérimentaux et pour l'annotation des images du jeu de données conçu.

Je souhaite exprimer ma gratitude envers tous mes amis et les rencontres que j'ai eu l'occasion de faire au cours de ces dernières années. Les moments de partage ont été des sources de bonheur inestimables. Les activités sportives avec Fan et Julien, mais également avec Laetitia, resteront gravées dans ma mémoire longtemps.

Enfin, je voudrais clore ces remerciements en exprimant ma profonde gratitude envers ma famille. Leur soutien indéfectible tout au long de mes années d'études, combiné à l'aide précieuse des enseignants qui ont jalonné mon parcours jusqu'à ce moment précis, ont été des piliers essentiels de ma réussite.

Table des matières

Table des figures

Liste des Abréviations

OMS	Organisation Mondiale de la Santé
AIPC	Agence Internationale pour la Prévention de la Cécité
CIM	Classification Internationale des Maladies
DMLA	Dégénérescence Maculaire Liée à l'Âge
DTI	Différence Temporelle Interaurale
DII	Différence d'Intensité Interaurale
HRTF	Head Related Transfer Function
LIDAR	Laser Imaging Detection And Ranging
IA	Intelligence Artificielle
ML	Machine Learning
ANN	Artificial Neural Network
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
EOA	Electronic Orientation Assistance
ETA	Electronic Travel Assistance
PLD	Position Location Devices
GPS	Global Positioning System
GSM	Global System for Mobile Communications
SVA	Substitution Visuo-Auditive
IMU	Unité de Mesure Inertielle
OSM	OpenStreetMap
GPU	Graphics Processing Unit
ARM	Advanced RISC Machines
FPGA	Field-Programmable Gate Array
FP32	Virgule Flottante sur 32 bits

1 Introduction

Sommaire

1.1 Contexte sociétal et enjeux de l'assistance aux personnes aveugles

L'absence d'acuité visuelle, aussi bien à la naissance que plus tardivement, affecte des millions de personnes à travers le monde. Les causes de la cécité sont nombreuses et variées, incluant des maladies génétiques, des infections, des traumatismes et des conditions liées à l'âge comme la dégénérescence maculaire. Ces multiples causes ont des conséquences dramatiques sur la qualité de vie, l'autonomie et l'insertion dans la société civile à travers l'éducation, l'emploi et la participation à des activités sociales.

1.1.1 Causes et conséquences sociétales croissantes

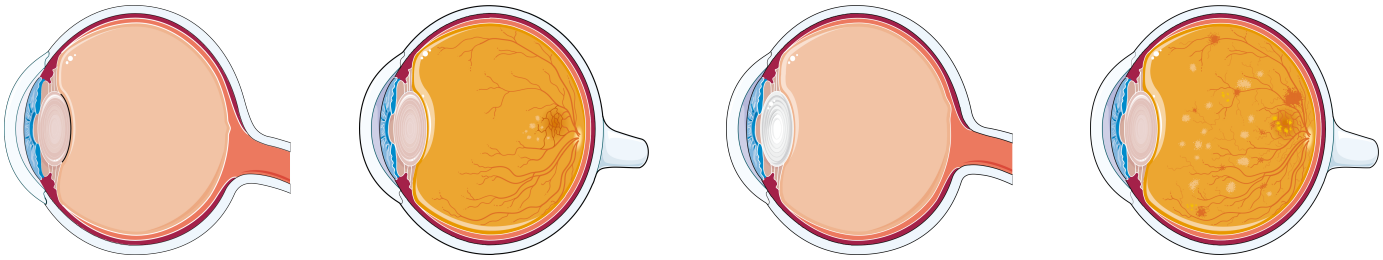
Depuis son lancement en 1999 par l'Organisation Mondiale de la Santé (OMS) et de l'Agence Internationale pour la Prévention de la Cécité (AIPC), l'initiative Vision 2020 : The Right to Sight s'est ardûment engagée à éliminer les principales causes évitables de cécité. Son objectif principal vise à alerter à la fois les gouvernements et le grand public sur les risques et les impacts des troubles visuels. Selon l'AIPC, en 2020, le nombre de personnes touchées par une déficience visuelle modérée à sévère dépassait les 295 millions, tandis que 43,3 millions étaient totalement privées de la vue [1]. La gradation de ces déficiences est codifiée dans la Classification Internationale des Maladies (CIM, ou ICD en anglais) et se base sur l'acuité visuelle du meilleur œil après correction [2]. Selon cette classification retranscrite à Table 1.1, une personne est considérée comme ayant une déficience visuelle sévère si son acuité est comprise entre 1/60 et 3/60. Cela signifie qu'elle perçoit à une distance d'1 à 3 pieds ($\approx$ 0.3 et 0.9 mètre) ce qu'une personne à la vue normale peut percevoir à 60 pieds ($\approx$ 18 mètres). L'acuité visuelle est mesurée à l'aide d'échelles ophtalmologiques spécifiques. Parmi celles-ci, l'échelle de Landolt, est caractérisée par une série d'anneaux brisés, et l'échelle de Monoyer, utilise l'alphabet latin. Néanmoins, l'omission de certains critères et l'évolution croissante des exigences en matière d'acuité visuelle dans notre société contemporaine suggèrent la nécessité de revoir cette classification [3].

Malheureusement, ces chiffres sont sur une trajectoire ascendante. D'ici à 2050, les estimations projettent que ces nombres pourraient grimper à 474 millions pour les déficiences visuelles modérées à sévères et à 61 millions pour les cas d'aveuglement total [4]. De nombreux facteurs sont à l'origine de cette augmentation préoccupante, en

Classification	Acuité visuelle		Champ central
	Maximum	Minimum ou égal à	
0 - Légère déficience visuelle	-	6 / 18	-
1 - Déficience visuelle modérée	6 / 18	6/60	-
2 - Sévère déficience visuelle	6/60	3/60	-
3 - Aveugle	3/60	1/60	$< 10°$ et $> 5°$
4 - Aveugle	1/60	Perception lumineuse	$\leq 5°$
5 - Aveugle	Pas de perception lumineuse		-
9 - Non spécifié	Non déterminé ou spécifié		

TABLE 1.1 – Catégories de sévérité du déficit visuel selon la Classification Internationale des Maladies [2].

tête de liste de laquelle on retrouve des affections telles que la dégénérescence maculaire liée à l'âge (DMLA) et les cataractes. La DMLA touche spécifiquement la macula, une petite région centrale de la rétine, provoquant une perte de la vision centrale, tandis que les cataractes résultent de l'opacification du cristallin de l'œil, engendrant une vision altérée, floue ou réduite. La démographie mondiale joue également un rôle dans cette croissance. En effet, d'après les projections démographiques reportées dans la figure 1.2, plus de la moitié (55%) de la population mondiale aura plus de 50 ans [5]. Cette tendance au vieillissement est particulièrement prononcée en Asie et en Europe, régions où la natalité diminue et l'espérance de vie augmente continuellement.



a. Œil sain **b.** DMLA **c.** Cataracte **d.** Rétinop. diab.

FIGURE 1.1 – Schéma d'un œil et de différentes causes de cécité.

Mais au-delà de la DMLA et des cataractes, d'autres affections médicales pourraient également intensifier le problème. Le diabète, par exemple, est une cause majeure de rétinopathie diabétique, une affection qui peut conduire à la cécité. De même, le glaucome, une maladie oculaire souvent liée à une pression intraoculaire élevée, continue de représenter une menace majeure pour la vision. La figure 1.1 illustre l'impact de diverses

affections oculaires sur un œil sain (a), notamment la dégénérescence de la macula associée à la DMLA (b), la cataracte caractérisée par l'opacification du cristallin (c), ainsi que les hémorragies intrarétiniennes et les œdèmes maculaires caractéristiques des rétinopathies diabétiques (d).

Bien que des traitements existent, comme la chirurgie, le laser ou les collyres pour le glaucome, et permettent de ralentir ou de limiter la progression de certaines maladies, une détection précoce de ces problèmes est primordiale pour réduire leur gravité et, par conséquent, leur impact sur la vie de millions de personnes dans le monde. Cependant, l'accès à des moyens de prévention, de détection et de traitement n'est pas équivalent pour toutes les populations mondiales. Il est ainsi limité dans certains pays en développement, ce qui expose ces populations à un risque accru de devenir totalement ou partiellement aveugles. Il est estimé que 90% des cas de perte visuelle sévère auraient pu être évités. La mise à disposition de ressources pour lutter contre l'apparition de ces pathologies est un enjeu mondial dans la lutte contre le développement de la cécité. Néanmoins, malgré les progrès dans le domaine médical, la médecine moderne ne peut pas encore restaurer totalement ou même partiellement la vue d'une personne aveugle. L'amélioration de la vie quotidienne de ces personnes et leur intégration dans nos sociétés constituent un défi majeur.

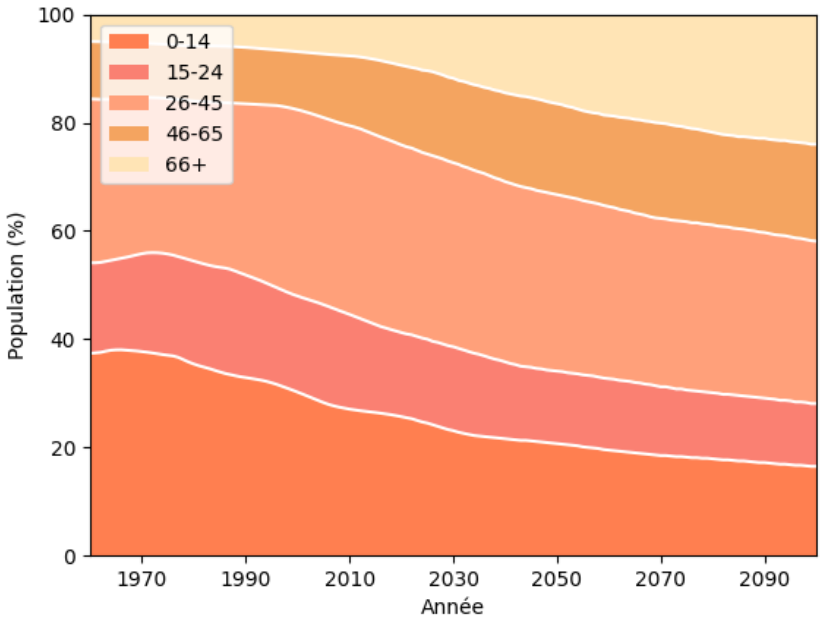


FIGURE 1.2 – Estimation et projection de l'évolution des groupes d'âge de 1960 à 2100 [5].

1.1.2 Intégration et impact dans la société

Au fil de l'histoire, la perception et l'intégration des personnes aveugles ont connu de profondes mutations. Dans l'Antiquité grecque et égyptienne, en l'absence d'explications scientifiques, la cécité était souvent perçue comme une punition divine pour un comportement déviant. Ainsi, Erymanthos fut rendu aveugle par Aphrodite après l'avoir vue, par mégarde, nue se baigner dans son bain. La cécité, souvent issue de guerres, d'accidents ou de maladies aujourd'hui éradiquées, était fréquente. Thucydide, par exemple, relatait en 429 av. J.-C. une épidémie de peste à Athènes, laissant les survivants aveugles [6]. À Rome, une certaine forme d'inclusion existait, les aveugles pouvaient jouir de certaines libertés et accéder à des hautes professions comme la magistrature, bien qu'il n'existe cependant aucun indice de l'accession de ces personnes à ces distinctions avant ou après être devenues aveugles [7]. Le Moyen Âge stigmatisa les personnes aveugles en les identifiant comme un groupe de personnes souvent assimilées à des mendiants dépravés, oisifs, voire menteurs sur leur réelle cécité. Ils étaient souvent regroupés au sein de communauté religieuse, vouée à la prière et à des tâches ingrates, telles que guider les condamnés à mort ou mendier pour la subsistance de leur groupe. De plus, cette période est synonyme de moquerie sur leur statut social et leurs maladresses [8]. Ce n'est qu'à l'époque moderne que la société a commencé à reconnaître et à représenter leur individualité, notamment dans des œuvres artistiques. Des philanthropes et des humanistes ont suggéré de les intégrer pleinement à la société en leur confiant des responsabilités, telles que des travaux manuels ou des fonctions de musicien. L'Âge des Lumières marqua un changement avec une approche pédagogique adaptée à leur handicap. Diderot affirma, dans son œuvre *Lettre sur les aveugles à l'usage de ceux qui voient*, que l'apprentissage tactile stimule des mécanismes mentaux distincts que ceux sollicités par la vision. Cette période historique inscrivit l'apparition des premières écoles dédiées aux personnes aveugles dans les grandes villes européennes, portées par des figures telles que Valentin Haüy, fondateur de l'Institution Nationale des Jeunes Aveugles à Paris.

Depuis cette époque, de remarquables progrès en psychologie cognitive et en médecine ont favorisé une intégration accrue des personnes aveugles dans la société. Ces avancées ont été soutenues par des méthodologies d'enseignement spécifiquement conçues pour répondre à leurs besoins. Toutefois, malgré ces améliorations, leur intégration demeure, dans certains cas, restreinte ou marginale par rapport à l'ensemble de la population. L'éducation pour les aveugles s'organise principalement au sein d'établissements spécialisés. Ces structures, dotées de professionnels formés pour répondre à

leurs besoins spécifiques, offrent un accès à des ressources dédiées. Cependant, plusieurs études ont démontré que leur intégration dans les écoles traditionnelles offre non seulement des avantages académiques, mais aussi des compétences particulières et des bénéfices sociaux. Ces avantages s'observent aussi bien chez les élèves aveugles que chez leurs camarades voyants. Cependant, une telle intégration nécessite une formation et un soutien renforcés pour les enseignants [9]. Par ailleurs, les niveaux de satisfaction et les sentiments exprimés au quotidien par les élèves aveugles sont comparables dans le système éducatif spécialisé ou traditionnel [10].

Le monde professionnel n'est pas non plus épargné par une exclusion des personnes malvoyantes où leurs taux d'emploi est significativement inférieur (10%) à celui de la population générale (70%) [11], et elles sont proportionnellement plus nombreuses à occuper des postes peu qualifiés (10% contre 5% dans la population générale). Mais les chiffres des niveaux de qualification sont peut-être erronés par le fait que ces métiers sont ceux où les risques d'accidents sont les plus élevés [12]. De plus, la non-intégration professionnelle des personnes malvoyantes est une préoccupation économique pour de nombreux États, en raison d'une perte de productivité associée, estimée à 411 milliards de dollars en 2020 [13].

1.1.3 Une qualité de vie et une autonomie dégradée

L'intégration insuffisante des personnes aveugles dans la société aggrave les multiples défis qu'elles rencontrent déjà au quotidien. En effet, de simples actions, considérées comme banales pour les individus voyants, deviennent des obstacles majeurs pour elles. L'incapacité à percevoir visuellement leur entourage affecte grandement leurs capacités à comprendre et par conséquent à interagir ou à se déplacer dans l'espace. Pour une personne malvoyante, le simple fait de se déplacer dans un environnement donné peut s'avérer extrêmement ardu. De plus, notre monde est majoritairement façonné pour les individus voyants où l'agencement du mobilier, qu'il s'agisse de lieux publics, d'espaces de travail ou même d'habitations privées, est souvent pensé en priorité pour son esthétisme et sa fonctionnalité pour les voyants, au détriment des contraintes des personnes aveugles ou malvoyantes. Cette conception, non inclusive, se manifeste dans la diversité des agencements et des configurations où chaque espace représente un défi distinct et nécessite de s'adapter, et de se familiariser. L'impact d'une mobilité dégradée se répercute dans leur accès à des activités essentielles comme le travail, l'éducation ou les interactions sociales en accentuant leur faible intégration dans la société. De plus, lors des heures de pointe,

la surcharge cérébrale liée aux flux importants peut être accablante, incitant certains à éviter totalement ces périodes. L'enseignement supérieur est un exemple de condensés d'obstacles pour les personnes malvoyantes, qui doivent coexister dans un campus bondé, où des foules d'étudiants se déplacent simultanément. De plus, les changements imprévus de salles de classe ou des itinéraires fluctuants qui nécessitent une adaptation constante. Pour de nombreux étudiants aveugles ou malvoyants, cette charge mentale est parfois si intense qu'elle peut les décourager jusqu'à les pousser à l'abandon de leurs études supérieures [14].

Des personnes appelées guide humains ou guides pour personnes malvoyantes sont formées pour accompagner et assister au quotidien les personnes aveugles. Ils sont les yeux de substitution qui permettent de participer pleinement aux événements sociaux et éducatifs dans des milieux non familiers. Cependant, bien qu'essentiels, ils peuvent aussi introduire une certaine dépendance, entravant ainsi la quête d'autonomie de la personne aveugle. D'autre part, les contraintes comme le nombre limité de guides disponibles et les coûts liés à leur service restreignent leur adoption comme support quotidien. Dans cette optique, des solutions qui n'impliquent pas l'assistance continue d'un tiers sont souhaitables comme l'utilisation accrue d'autres modalités sensorielles non perturbées par un handicap comme le toucher ou l'ouïe.

1.2 Perception Multisensorielle

1.2.1 Vision : Un rôle majeur

Les sens humains sont nos capteurs physiologiques qui nous permettent d'appréhender notre position au sein de l'environnement qui nous entoure. Les informations des différentes modalités sensorielles sont agrégées ensemble pour fournir à l'être humain une connaissance élargie afin de favoriser un comportement ou un sentiment de confort, de sécurité. Ces mécanismes sont par exemple utilisés sur des consommateurs par des méthodes de marketing sensoriel pour déclencher un achat avec l'effusion d'odeurs combinées à la diffusion de mélodie agréable [15]. Néanmoins, bien que toutes les informations sensorielles soient utilisées pour représenter notre environnement, leur impact varie en fonction de la tâche souhaitée. La vision joue un rôle prédominant parmi les autres sens en termes de capacité à recueillir une vaste quantité d'informations en un coup d'œil. En effet, la vision possède une résolution spatiale nettement supérieure à celle des autres sens,

rendant son utilisation plus facile dans des tâches telles que la reconnaissance spatiale, l'identification d'objets ou d'éléments dans notre entourage [16], [17]. Un exemple illustratif de cette prédominance visuelle se trouve dans la navigation en milieu obscur. Dans un tel contexte, les êtres humains ont tendance à privilégier les endroits qu'ils ont déjà vus, plutôt que de compter davantage sur leurs autres sens pour se repérer et interpréter leur environnement [18]. Par ailleurs, la vision sert de référence [19] et de calibration durant l'enfance pour les autres sens [20]. Toutefois, cette tendance pourrait diminuer avec le temps [21]. Une telle prédominance visuelle pourrait en réalité compenser une lacune intrinsèque du système visuel humain face à des événements imprévus ou soudains, contrairement à nos systèmes auditifs et tactiles [22]. La domination de la vision présente cependant des limites lorsque les informations visuelles sont manquantes dues par exemple à une cécité visuelle. En effet, en l'absence de capacité visuelle, l'acquisition d'une connaissance spatiale essentielle de l'environnement est entravée. L'accès à une cartographie mentale devient alors limitée, reposant principalement sur des informations transmises par d'autres individus ou sur celles perçues grâce aux autres sens.

1.2.2 Compensation sensorielle

L'être humain aveugle utilise ainsi différents canaux sensoriels pour percevoir son espace. Cette aptitude est augmentée par la capacité adaptative intrinsèque du cerveau humain à subir des réorganisations structurelles, connue sous le nom de plasticité cérébrale, ce qui représente un mécanisme fondamental dans le processus de compensation sensorielle. Cette plasticité se manifeste principalement par la création de nouvelles connexions neuronales et par le renforcement de connexions préexistantes [23]. La figure 1.3 décrit le phénomène d'apparition et de réduction des connexions synaptiques entre neurones. Cette adaptabilité cérébrale est influencée par une interaction complexe de facteurs génétiques et environnementaux.

De récentes recherches suggèrent que suite à la défaillance d'un sens, les aires cérébrales associées à ce dernier peuvent subir une reconversion afin de traiter les informations issues d'autres modalités sensorielles. Chez les individus malvoyants, l'aire corticale visuelle pourrait être réaffectée à des taches auditives ou tactiles [24]. L'alliance de cette adaptabilité naturelle du cerveau et d'un entraînement intensif permet aux personnes atteintes de déficience visuelle d'accroître leur sensibilité aux stimulus auditifs et haptiques. Cette combinaison leur offre une compréhension plus nuancée et enrichie de leur environnement. Toutefois, la réduction graduelle du nombre de connexions synap-

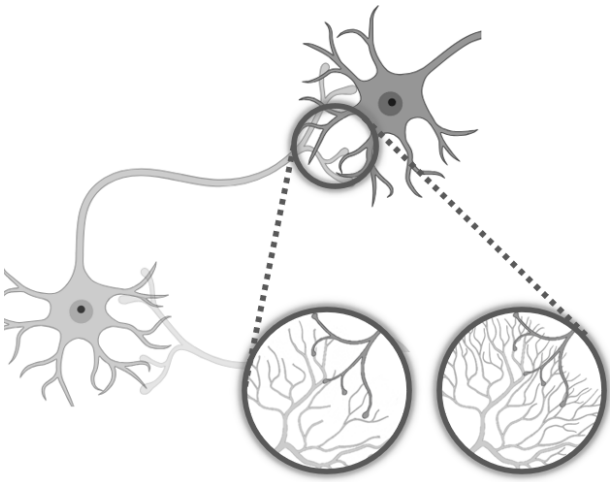


FIGURE 1.3 – Schéma de l'évolution des connexions neuronal.

tiques au fil des ans limite la plasticité naturelle. Cette réduction restreint le potentiel de compensation, surtout en cas de cécité survenue à un âge avancé [25]. Des performances comportementales supérieures et des aptitudes neuroplastiques intermodaux ont été remarqués chez les aveugles en bas âge. Une période dite critique de 14 ans [26] à 16 ans [27] au-delà duquel la plasticité cérébrale devient moins performante, mais conserve un niveau élevé [28]. De plus, des techniques d'intervention modernes, comme la neurostimulation, peuvent augmenter cette plasticité, améliorant les capacités perceptuelles dans d'autres domaines sensoriels. Malheureusement, les éléments non perceptibles par les autres modalités sensorielles ne peuvent se substituer aux informations visuelles par un mécanisme de compensation sensorielle.

1.2.3 Assistance à la mobilité

L'accès à des informations visuelles distantes et non accessibles constitue un enjeu majeur pour garantir une navigation autonome et sécurisée au sein de tout espace. Au fil de l'histoire, diverses aides à la mobilité ont vu le jour afin d'amplifier la perception sensorielle des éléments présents dans un environnement. Historiquement, les interactions entre l'homme et l'animal ont initié une révolution dans la manière d'appréhender le monde pour les individus malvoyants. Les chiens d'assistance, plus communément appelées **chien-guide**, se sont imposés comme une solution viable et efficace pour guider leurs maîtres. Après une longue et coûteuse formation, leurs sens aiguisés et leurs compétences permettent, en plus d'offrir de simples indications directionnelles, une analyse et une anticipation d'un éventuel danger et obstacle. De son côté, la **canne blanche**

s'est imposée comme un outil indispensable. Elle agit comme un prolongement du corps, offrant une compréhension immédiate sur l'environnement proche grâce au retour sonore et tactile qu'elle génère lorsqu'elle entre en contact avec un obstacle. Cependant, malgré leur efficacité, ces aides sont restreintes, que ce soit en termes de portée, de profondeur de l'information transmise, ou même du coût financier et du temps de formation nécessaire. Face à ces défis, les avancées scientifiques et technologiques récentes ont donné naissance à de nouvelles solutions. Les dispositifs d'assistance modernes, combinant les principes de science cognitive, des capteurs de pointe et des algorithmes de traitement d'information sophistiqués, agissent comme des interfaces dont l'objectif est de transformer les informations du monde extérieur en données accessibles pour l'utilisateur. Les interactions novatrices basées sur des encodages d'informations visuelles par des signaux nerveux, sonores ou tactiles précis offrent une expérience multisensorielle enrichissante.

1.3 Plan de thèse

Cette thèse expose le développement d'un système d'aide à la mobilité fondé sur une substitution sensorielle des informations visuelles en signaux sonores spatialisés. Le manuscrit débute par le chapitre 2 avec un état de l'art des méthodes d'assistances existantes et des mécanismes propres aux méthodes de substitution sensorielle que sont respectivement les processus d'encodage sonore et d'extraction d'information pertinente par des mécanismes de vision artificielle. Puis, les chapitres 3 et 4 s'intéresseront à nos contributions sur ces deux axes. Finalement, le chapitre 5 détaille l'intégration de ces composantes dans un système d'assistance autonome complet pour la mobilité.

Le chapitre 2 retrace l'évolution des méthodes d'assistance pour les personnes malvoyantes, traditionnellement appuyées par l'aide humaine, mais souffrant de lacunes préexistantes. Puis, il expose des innovations technologiques visant à pallier ces lacunes, notamment à travers des dispositifs de substitution sensorielle qui promettent d'améliorer la perception de l'information par des moyens artificiels et autonomes. L'accent est mis sur les techniques convertissant les données visuelles en signaux haptiques ou auditifs, offrant ainsi des informations additionnelles utiles, bien que leurs efficacités varient selon le protocole d'encodage utilisé. Le chapitre détaille ces méthodes d'encodage, leurs impacts sur l'interprétation des données, puis les méthodes de vision artificielle permettant de définir la présence d'éléments pertinents à transmettre à l'utilisateur. Une revue des méthodes, en particulier d'intelligence artificielle, permet une analyse détaillée des données visuelles. Ces approches visent à procurer des détails plus fins de l'espace, facilitant ainsi

une compréhension plus approfondie par l'utilisateur.

Le chapitre 3 introduit une méthode d'encodage d'informations spatiales utilisant un son spatialisé unique pour représenter la localisation d'un élément ou d'un ensemble de positions décrivant les obstacles à proximité d'une personne. Une évaluation de cette représentation sonore 3D est proposée afin de mesurer la précision de localisation selon les différentes dimensions spatiales. Ensuite, le chapitre expose l'intégration de notre méthode d'encodage auditif au cœur d'un dispositif conçu pour guider méthodiquement un utilisateur malvoyant vers une destination spécifique à l'intérieur d'un édifice. Une expérimentation est conduite pour évaluer l'efficacité de cette méthode d'encodage pour l'interprétation de la trajectoire à suivre et de la position des obstacles.

Le chapitre 4 présente une approche intégrant des méthodes de vision artificielle pour une analyse fine d'un espace de navigation urbain composé d'obstacles et de zones dangereuses pour un piéton. Cette approche principalement construite autour d'algorithmes par apprentissage profond est décrite parcourant la création d'une nouvelle base de donnée définie pour l'assistance aux personnes piétonnes, les limitations de ces méthodes et les ouvertures possibles.

Enfin, le chapitre 5 illustre la synthèse des avancées en matière d'encodage de l'information visuelle en signaux sonores spatialisés et d'analyse visuelle d'environnements urbains par le développement d'un prototype de système d'assistance à la navigation urbaine autonome adaptée aux contraintes des personnes malvoyantes. Ce système, caractérisé par une consommation énergétique modérée, est structuré par une parallélisation des différents processus impliqués jusqu'à la génération des stimulus sonores. Une série d'expérimentations menées dans divers contextes urbains, tels que des zones piétonnes ou des espaces partagés, est détaillée, mettant en lumière les performances et les adaptations du système dans le monde réel.

2

État de l'art

Sommaire

2.1 Méthodes et solutions d'assistance destinées aux personnes malvoyantes

L'assistance aux personnes aveugles ou malvoyantes s'est manifestée comme une préoccupation fondamentale à travers les âges avec l'ambition de forger une société empreinte d'inclusion et dénuée de discrimination. Au cours de l'Antiquité, cette catégorie de la population était souvent reléguée à un statut inférieur, une notion illustrée par Quintilien dans son œuvre *De institutione oratoria, livre VIII*. Il y évoque, par la formule *"On dit d'un aveugle qu'il se tient au deçà de la vie"*, l'idée que les malvoyants vivaient en marge de l'existence pleine et entière. Le rapport aux personnes déficientes évolua au cours du temps à la recherche d'une meilleure inclusion dans la société, mais également d'une plus grande autonomie pour se déplacer, accéder à des services primordiaux. Face à l'impératif d'intégrer ces personnes dans le tissu sociétal, politique et économique comme des citoyens à part entière, des initiatives innovantes ont vu le jour à travers l'histoire. Elles ont pour vocation d'augmenter la compréhension de l'environnement, en leur offrant une perception élargie des éléments qui leur sont biologiquement inaccessibles. La figure 2.1 illustre l'évolution des méthodes et des systèmes d'assistance, en commençant par l'entraide interhumaine, la forme la plus instinctive d'assistance, pour finir avec les innovations technologiques actuelles. Cette chronologie met en lumière des étapes clés qui ont constitué des tournants majeurs dans la manière dont les individus déficients

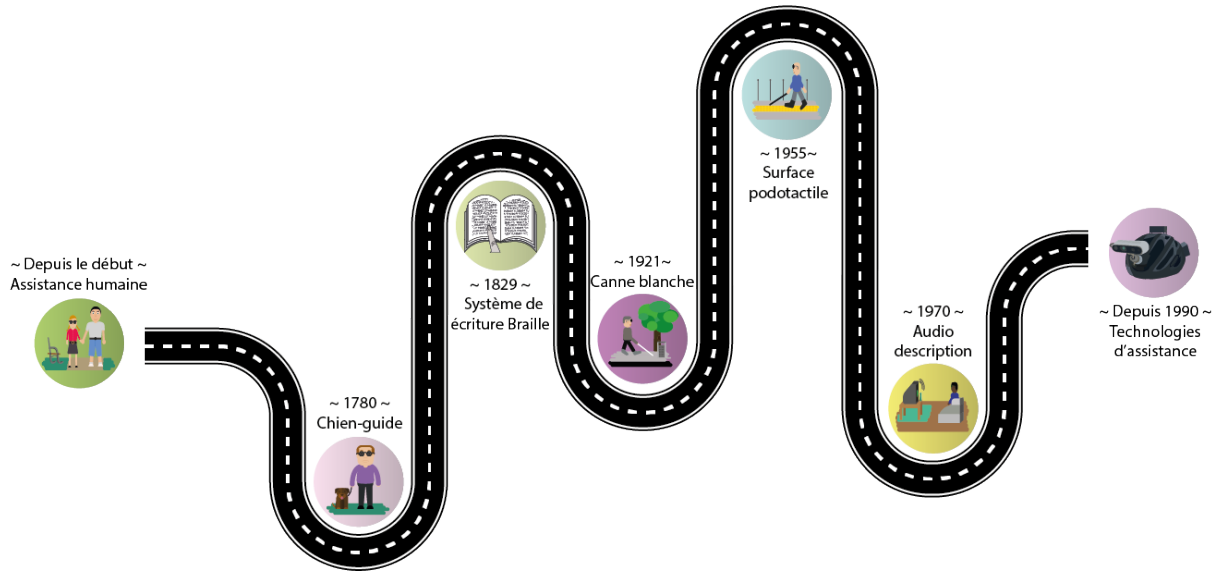


FIGURE 2.1 – Évolution des solutions d'assistance aux personnes aveugles au cours de l'histoire.

visuels appréhendent et interagissent avec leur environnement.

2.1.1 Méthodes traditionnelles

L'une des premières solutions proposées aux personnes aveugles, outre l'interaction avec les personnes voyantes, a été l'utilisation d'animaux entraînés et en particulier les chien-guides dans le rôle d'auxiliaires. L'origine de leur emploi remonte à l'époque romaine et simultanément en Asie comme en témoignent des fresques. Cependant, leur usage, s'est démocratisé à partir de la fin du $XXIII^{\text{ème}}$ siècle où, au sein de l'hôpital des Quatre-Temps (Paris), le premier essai structuré de dressage de chiens pour les aveugles a vu le jour et s'est ensuite propagé à travers l'Europe. La Première Guerre mondiale marqua un tournant majeur pour les chien-guides liés à l'augmentation tragique du nombre de soldats aveuglés par les gaz toxiques ou les explosions d'obus. Afin de faire face à ce besoin croissant, le Dr Gerhard Stalling décida de fonder une école de dressage pour former, en masse, des chiens capables d'aider ces vétérans blessés dans leur quotidien. Ces chiens ont pour objectif de renforcer l'autonomie des personnes malvoyantes lors de leurs déplacements. En effet, ils leur permettent un déplacement libre, d'éviter les obstacles et d'interagir avec leur environnement [29]. La vision des chiens se substitue ainsi à celle de leur propriétaire. Cette relation repose sur une confiance mutuelle où une erreur, comme un événement soudain mal jugé lors d'un passage piéton par exemple, peut ébranler cette confiance et augmenter le stress du propriétaire. Mais au-delà de la fonction de navigation des chien-guides, ils apportent un soutien émotionnel immense. Pour beaucoup de propriétaires, ces chiens ne sont pas seulement des outils, mais des membres à part entière de leur famille. Le lien profond entre un individu aveugle et son chien-guide est si fort que la perte ou la séparation peut causer une profonde dépression. En outre, la présence d'un chien-guide est un repère clairement identifié par les personnes voyantes de la présence d'un individu malvoyant, facilitant ainsi les interactions et l'assistance, notamment dans des situations délicates ou lorsqu'il s'agit d'accéder à certaines informations.

Néanmoins, l'intégration des chiens guides dans la société rencontre des défis. Leur formation est longue, complexe et s'accompagne de coûts élevés [29]. De plus, certaines personnes, par peur ou ignorance, peuvent être réticentes ou discriminantes envers ces animaux, surtout si elles sont affectées par la cynophobie. Dans certains établissements, comme les restaurants, les personnes aveugles et leurs chien-guides peuvent se heurter à des refus d'accès, malgré le rôle vital de ces animaux pour leurs propriétaires. Les

transports en commun, essentiels pour beaucoup de personnes aveugles, peuvent aussi poser un problème. Des récits de discrimination ou d'incompréhension, allant du simple refus d'accès au manque de place adaptée pour le chien, sont malheureusement courants. Les taxis, mais également le secteur aérien représentent un exemple important de difficulté d'accès. En effet, un nombre conséquent de personnes aveugles ont rapporté l'absence d'options claires pour voyager avec un chien-guide [30] dès l'enregistrement en ligne. De surcroit, une fois à l'aéroport, le manque d'aménagements spécifiques et d'assistance adaptée rendent le déplacement stressant et compliqué tandis avec l'avion, la situation n'est guère meilleure avec parfois un manque d'espace pour le chien et un manque d'information du personnel de bord. Ces barrières récurrentes dissuadent certaines personnes aveugles de demander l'assistance d'un chien-guide [29] ou plus globalement de voyager seul.

FIGURE 2.2 – Canne blanche.

La canne blanche, représentée dans la figure 2.2 demeure, avec le chien d'aveugle, la solution la plus répandue et plébiscitée au quotidien pour le déplacement des personnes aveugles. Introduite au début du $XX^{\text{ème}}$ siècle, la canne blanche est une réponse à la fois économique et ingénieuse aux besoins des personnes aveugles. Fabriquée à partir de matériaux légers comme le métal ou la fibre de carbone, elle bénéficie fréquemment d'une conception rétractable, optimisant ainsi son rangement et son transport. Elle sert d'extension aux bras de son utilisateur, permettant la détection d'obstacles, facilitant le franchissement de portes ou d'escaliers grâce à un retour sensoriel, qu'il soit haptique ou auditif. Les vibrations ressenties via le manche de la canne peuvent renseigner sur la nature du sol, une variation de hauteur ou encore sur la constitution d'un obstacle. Les sons produits lors de la frappe avec différents objets fournissent également des indications précieuses, chaque matériau produisant une sonorité distincte. De plus, la canne blanche

couplée avec une surface podotactile, illustrée à la figure 2.3, peut s'avérer être très efficace lors des déplacements. En effet, ces revêtements podotactiles, inventés en 1980 au Japon avec leurs plots distinctifs en relief, offrent un guide tactile au sol. Chaque plot, qu'il soit en forme de bande ou de cylindre, est soigneusement conçu pour être ressenti à travers la plante des pieds et la canne blanche. Les plots sont disposés selon des grilles régulières pour signaler des dangers comme le début d'un escalier, l'approche d'une intersection ou la lisière d'un quai de gare. La teinte marquée et distincte du revêtement offre une reconnaissance aisée pour ceux ayant une vision réduite. Néanmoins, la variété des motifs et leur disposition peuvent parfois désorienter les personnes entièrement aveugles. De surcroît, les reliefs podotactiles, bien que cruciaux pour les malvoyants, peuvent constituer des entraves pour les personnes à mobilité réduite circulant en fauteuil roulant. En particulier, les fauteuils sans systèmes d'amortissement peuvent subir des secousses notables en traversant ces zones, occasionnant un inconfort pour l'utilisateur. L'établissement d'une norme unifiée, associée à un code couleur standardisé, faciliterait grandement l'orientation des personnes aveugles, tout en rendant la formation à l'utilisation de ces repères plus intuitive [31]. La pandémie de Covid-19 en 2019 a révélé les lacunes des systèmes podotactiles. En effet, avec l'implémentation de sens de circulation dans de nombreux espaces publics, de nombreuses personnes aveugles se sont retrouvées désavantagées, privées d'informations directionnelles liées à la distanciation sociale. Toutefois, l'introduction de motifs podotactiles unidirectionnels a été suggérée comme une solution pour pallier ce manque [32].

FIGURE 2.3 – Surface podotactile.

2.1.2 Technologies d'assistance

Depuis plusieurs décennies, l'évolution technologique a permis l'émergence de dispositifs innovants en complément des méthodes traditionnelles d'assistance aux personnes malvoyantes. Ces outils, regroupés sous l'appellation de *technologies d'assistance visuelle*, ont été spécialement conçus pour résoudre diverses problématiques rencontrées au quotidien par ces personnes, qu'il s'agisse de l'accès à des contenus écrits, de la navigation dans des lieux non familiers ou encore de l'orientation dans des espaces complexes. Sous ce terme, sont regroupés une myriade de méthodes différentes, invasives ou non. Une classification reportée dans la figure 2.4, suivant celle proposée ultérieurement [33], propose de subdiviser dans trois branches principales ces approches en fonction de leurs caractéristiques.

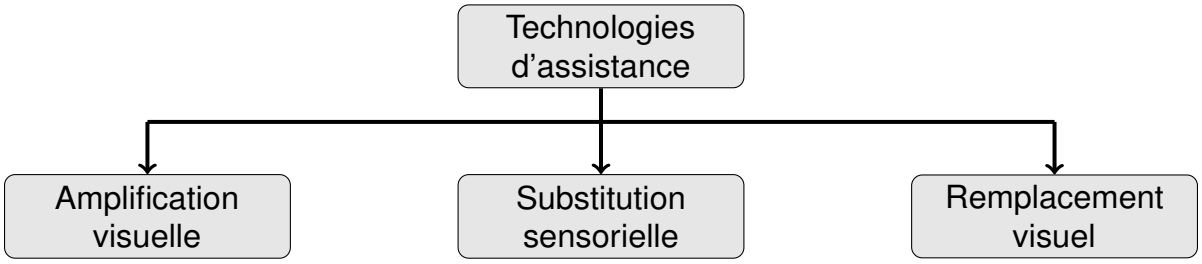


FIGURE 2.4 – Taxonomie des technologies d'assistance.

La catégorie de système dénommée *remplacement visuel* se différencie notablement des autres par sa fusion des méthodologies issues des sphères médicale et technologique. Cette catégorie de méthode est caractérisée par la stimulation directe, soit du nerf optique [34], soit du cortex visuel [35], ou encore de la rétine [36], [37], permettant de contourner la partie oculaire atteinte par une déficience. La figure 2.5 illustre le système visuel et les éléments précédemment cités. Une stimulation de ces zones par un signal électrique permet la manifestation d'un phénomène appelée phosphène, perçu comme un bref éclair. L'illusion perçue semblable à un élément visuel simple représente une étape significative dans la restitution partielle de la vision pour les individus malvoyants. Ces phosphènes sont générés à partir de dispositifs électroniques sophistiqués issues d'une information visuo-spatiale capturée par un système optique puis traitée et simplifiée. Le processus réalisé par une unité de traitement spécialisée telle qu'un microordinateur disposé sur un support ergonomique comme des lunettes électroniques avancées [38]. Ces informations sont ainsi transmises via une communication filaire ou non jusqu'à des matrices d'électrodes placées directement sur la zone souhaitée. Leur mise en place nécessite ainsi une chirurgie micro-invasive, durant laquelle ces électrodes et les systèmes

de communication sont implantés.

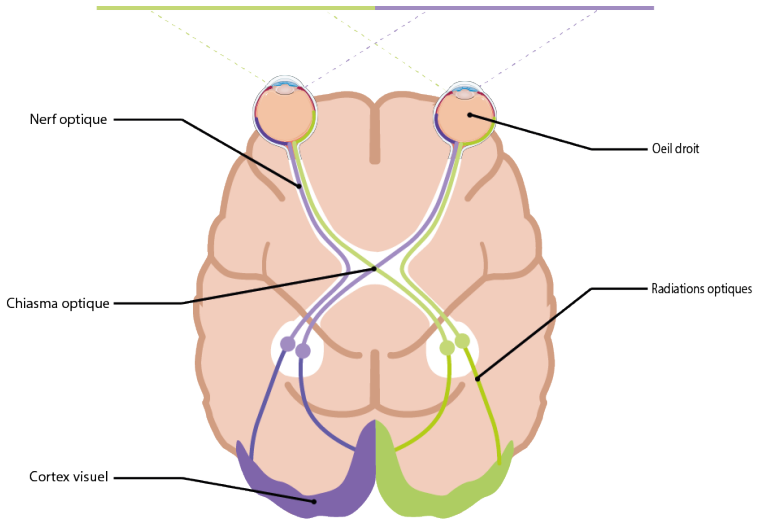


FIGURE 2.5 – Vue schématique du système visuel.

Le placement des électrodes à différentes positions du système visuel, comme le cortex ou la rétine, présente divers avantages. En effet, la stimulation directe de la rétine tire parti des mécanismes visuels complexes préexistants, tout en évitant les risques associés à une intervention directe sur le cortex. Cependant, l'implantation d'électrodes au niveau du cortex visuel permet d'émettre directement des stimulus liés à des zones spécifiques de l'espace visuel, en raison de la corrélation existant entre ces deux espaces [39]. Des expérimentations, menées initialement sur des animaux puis cliniquement sur des sujets humains, ont validé le potentiel de ces techniques pour la perception d'entités statiques, comme des symboles ou des lettres [40], mais également pour des séquences dynamiques grâce à l'usage de plusieurs électrodes. La densité des électrodes, représentés sous forme de patch ou de matrice d'électrodes sur le cortex visuelle, influence les capacités de perception. Par exemple, l'identification d'objets nécessaires exige ainsi l'implantation d'un patch d'émetteur, représenté sous forme de matrice d'électrodes de 16 par 16, tandis que la reconnaissance de formes nécessaires pour une navigation sûre dans un espace nécessite une matrice de dimensions moindres [41]. Des tâches plus complexes liées à la mobilité des personnes aveugles restent limitées [42].

Contrairement aux méthodes de remplacement visuel, qui visent à compenser une déficience visuelle à travers une exploitation des capacités de la rétine ou du cortex visuel, les approches de substitution sensorielle ou d'amplification visuelle proposent des solutions non invasives et temporaires pour pallier les déficiences visuelles chez les personnes malvoyantes. La loupe illustre parfaitement le moyen d'agrandir une information visuelle. Elle aide les gens à lire des documents sur lesquels la police d'écriture est trop réduite.

Certains dispositifs dérivés des longues-vues permettent d'accéder à des informations éloignées dans la rue comme les panneaux de signalisation ou des noms de rue. Des outils technologiques basés sur les possibilités offertes par les smartphones comme le GPS, l'appareil photo ou les cartes préenregistrées assistent [43] et permettent d'orienter les personnes déficientes dans un milieu urbain. Néanmoins, leur efficacité en extérieur est parfois compromise par des problématiques de reflets ou la difficulté de visualisation en plein soleil. Les casques de réalité augmentée avec leurs facultés d'enrichir une scène visuelle avec des éléments synthétiques, tels que le casque Hololens de Microsoft, sont également des outils intéressants pour désigner une orientation à suivre ou agrandir une information [44], [45].

2.2 Substitution sensorielle

Les méthodes de substitution sensorielle appartiennent à la dernière catégorie de méthode technologique d'assistance à destination des personnes aveugles. Ces technologies s'articulent autour de l'axiome de la transposition, convertissant des informations rendues imperceptibles en raison d'une défaillance d'une modalité sensorielle en signaux intelligibles pour une modalité intacte. Cette transposition s'appuie sur le mécanisme naturel de plasticité cérébrale, tirant avantage de la capacité singulière de redéfinir des nouveaux circuits neuronaux. Cette faculté adaptative intensifie les perceptions des modalités sensorielles préservées. Ces systèmes sont perçus avec un grand potentiel pour l'assistance aux personnes aveugles [46], [47] par leurs capacités intrinsèques en jouant avec les capacités naturelles humaines et leurs larges spectres de possibilités d'assistances quotidiennes. Ces méthodes permettent d'assister des personnes aveugles dans leur quotidien pour accéder à des informations visuelles telles qu'une retranscription d'un texte ou une assistance dans leurs mobilités. En effet, ces méthodes abordent plusieurs éléments nécessaires pour une navigation sécurisée dans des espaces non familiers, tels que la reconnaissance spatiale des objets environnants, la détermination de la trajectoire et la localisation de l'individu dans l'espace. Ces trois éléments, illustrés dans la figure 2.7, sont cités dans diverses catégorisations de méthodes et de systèmes comme suit :

- **Aide électronique aux déplacements** ou *Electronic Travel Aids (ETA) [48]* : Identifie et signale la position des obstacles ainsi que des zones dangereuses environnantes à la personne malvoyante, dans le but de faciliter la construction d'une représentation mentale de son environnement.
- **Aide électronique à l'orientation** ou *Electronic Orientation Aids (EOA)* [49] : Guide

la personne aveugle en lui indiquant le trajet optimal à suivre.

- **Dispositif de localisation** ou *Position Locator Devices (PLD)* : Détermine la position précise de la personne aveugle par rapport à son environnement.

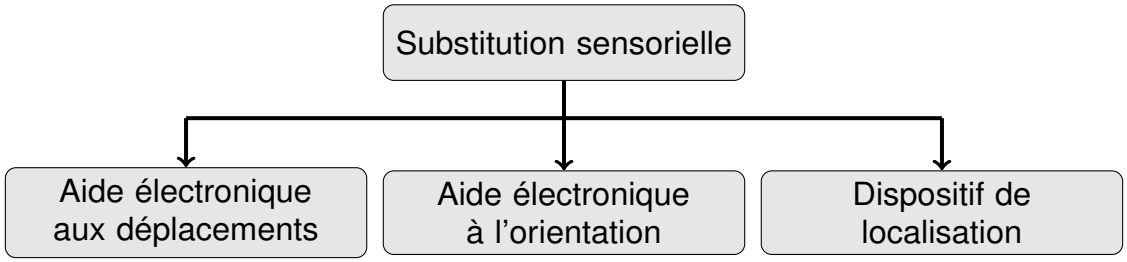


FIGURE 2.6 – Taxonomie des méthodes de substitution sensorielle dédiées à la navigation.

Le fonctionnement global de ces systèmes repose ainsi sur quatre composantes représentées dans le schéma de la figure 2.7. Une information spatiale issue de l'environnement est d'abord collectée et traitée afin d'isoler les éléments pertinents et effectuer une action spécifique. Ces informations sélectionnées subissent ensuite une transformation pour devenir un signal distinctif, intelligible par une modalité sensorielle alternative. Enfin, ce signal transcodé est transmis à l'utilisateur.

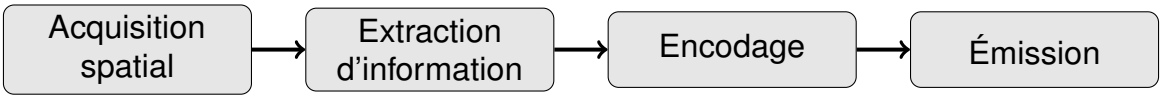


FIGURE 2.7 – Structure simplifiée d'un système de substitution sensorielle.

L'encodage des informations sensorielles vers une modalité de substitution, quelle que soit la finalité de ces méthodes, est un enjeu déterminant pour leur accomplissement. Un encodage imprécis ou inadapté à la situation peut avoir des conséquences importantes pour les personnes malvoyantes. Le choix d'un protocole d'encodage approprié, ainsi que plus globalement d'une modalité de substitution pertinente, s'avère essentiel. En effet, les capacités de substitution et de perception spatiale varient grandement entre deux sens humains et sont représentées dans le tableau 2.1 [50]. Ainsi, du fait de leurs propriétés fondamentales, l'ouïe et le toucher se positionnent comme les sens les plus adaptés pour pallier les déficiences visuelles. La perception haptique, influencée par les récepteurs cutanés, présente une capacité perceptive qui pourrait être comparable à une vision imparfaite ou floue [51]. De son côté, l'ouïe offre une perception qui se situe entre la vision et le toucher [52]. Cependant, les aptitudes cognitives de l'homme à traiter des informations simultanément demeurent limitées [53]. Une charge cognitive trop importante a un impact néfaste pour l'homme. Une priorisation de l'information critique aux détriments

des autres [54] est requise tout en considérant les problèmes propres à l'ensemble des systèmes d'assistances comme le confort de l'utilisateur, l'encombrement ou l'accessibilité.

Modalité sensorielle	Vue	Toucher	Ouïe	Odorat
Champ de perception	Large	Restreint	Large	Large
Information multiple	Oui	Champ de perception restreint	Interférence entre les signaux	Interférence entre les signaux
Détection d'objet	Précis	Précis sur de faible champ de perception	Moins précis que la vision	Moins précis que la vision
Identification d'objet	Précis	Moins précis que la vision	Moins précis que la vision	Très faible

TABLE 2.1 – Comparaison des informations perçues par les différents sens humain [50].

2.2.1 Substitution haptique

Une perception sensorielle liée au touché engage une combinaison de récepteurs de natures différentes afin d'interpréter la nature de l'objet. Ainsi, les mécanorécepteurs, un ensemble de terminaisons nerveuses au niveau de l'épiderme, renseignent sur la rugosité ou la texture d'un élément. Le système de lecture Braille s'appuie sur ces récepteurs pour interpréter la signification des motifs en relief. Les propriocepteurs, ancrés dans les tissus musculaires, transmettent au cerveau des informations concernant la localisation de nos membres. Les thermorécepteurs, pour leur part, offrent une perception sensorielle fine focalisée sur la température. L'ensemble de ces informations sensibles est transmis au cerveau, procurant au sens du toucher une perception diversifiée d'un élément, que la vue ou l'audition ne peuvent offrir. La répartition des récepteurs cutanés varie au niveau de l'épiderme ou des muscles. Certaines régions, telles que la surface palmaire ou les lèvres, présentent une densité accrue de ces récepteurs de l'ordre du millimètre [55], conférant ainsi une sensibilité supérieure à ces zones.

Les technologies de substitution haptiques s'appuient sur ces nuances pour substituer à une information visuelle une information compréhensible par une personne malvoyante à travers l'émission de stimulus artificiels générés à partir d'émetteurs disposés au contact du corps humain. Selon les exigences du contexte et le mode d'encodage sensoriel, ces stimulus peuvent être générés soit passivement, sans intervention de l'utili-

sateur, soit activement, nécessitant une action de l'utilisateur. Une substitution passive est privilégiée généralement lors d'une assistance à la mobilité en émettant promptement des alertes face à des dangers sans attendre une action humaine. Par exemple, des dispositifs électroniques, extensions modernes des cannes blanches traditionnelles, ont été conçus pour enrichir le champ perceptif des personnes malvoyantes. Le système *ELC* [56], émet une vibration au niveau du manche de la canne lorsqu'un obstacle est détecté. La distance entre l'utilisateur et l'obstacle est également fourni à l'utilisateur via des fréquences de vibrations supérieures lorsque l'objet est proche. De manière similaire, d'autres cannes électroniques adaptent l'intensité des stimulus en fonction de la distance séparant l'utilisateur de l'obstacle [57] bien que parfois, son interprétation soit mal perçue par les utilisateurs [58]. Les stimulus actifs peuvent cependant être associés en complément pour une interaction approfondie de l'environnement afin d'identifier un objet sans impératif de temps tout en conférant de meilleures capacités de reconnaissance de forme [59].

La proprioception, étymologiquement dérivée des termes latins "proprio" signifiant "soi-même", et "perception", englobe l'aptitude intrinsèque du corps humain à interpréter la localisation des muscles dans l'espace, mais également à percevoir la position d'un stimulus haptique. Cette aptitude de localisation d'un stimulus peut être associée avec l'utilisation de plusieurs répartis sur une surface, offre la possibilité de recréer un espace spatial dans le domaine haptique. Par exemple, *Path Force Feedback* [60] est une ceinture abdominale électronique équipée d'émetteurs répartis le long de celle-ci. L'émission d'un stimulus informe de la présence d'un obstacle proche, mais également son orientation azimutale relative. Des systèmes analogues recréent des espaces bi-dimensionnels à partir de surface d'émetteur comme *Tongue-placed electro-tactile* [61]. Ce dispositif vise à orienter une personne non-voyante au sein d'un environnement tout en indiquant la distance restante à parcourir à partir de la position d'un stimulus. Plus précisément, une information visuelle capturée par une caméra montée sur des lunettes électroniques est traitée par un ordinateur, puis transmise à des émetteurs situés sur la langue de l'individu. Ces émetteurs forment une configuration en étoile à huit branches de multiples tailles (figure 2.8) où la position sur le sommet de l'étoile et sa distance par rapport au centre indique respectivement l'orientation et la distance. Une technique statique transformant les niveaux de gris d'une image en signaux haptiques via une grille de dimensions 20×20 a démontré la capacité d'une personne à distinguer la forme d'un objet après une formation de 5 à 15 heures, ainsi que sa capacité à identifier la présence ou l'absence de lunettes sur un individu [62].

Les mécanismes thermorécepteurs de la peau humaine sont aptes à discerner les variations thermiques, telles que la proximité d'une source de chaleur ou de fraîcheur.

Certains dispositifs d'assistance à l'orientation intègrent des indicateurs thermiques situés sur le poignet d'une canne destinée aux malvoyants afin de signaler la présence d'un élément significatif dans leur environnement immédiat [63]. L'intégration de stimulus hétérogènes, qu'ils soient thermiques, vibratoires ou temporels, augmentent la finesse et la diversité des informations communiquées [64], [65]. Cette stratégie multimodale permet, après un apprentissage approfondi, d'identifier des objets par des combinaisons haptiques spécifiques. En se familiarisant avec ce langage tactile enrichi, les utilisateurs peuvent décoder avec précision la nature d'un objet ou d'un obstacle imminent, facilitant ainsi des interactions telles que la manipulation d'une porte ou l'évitement d'une menace potentielle. La pertinence des méthodes de substitution haptique s'est avérée probante dans des études portant tant sur la reconnaissance de formes que sur la localisation d'objets [66].

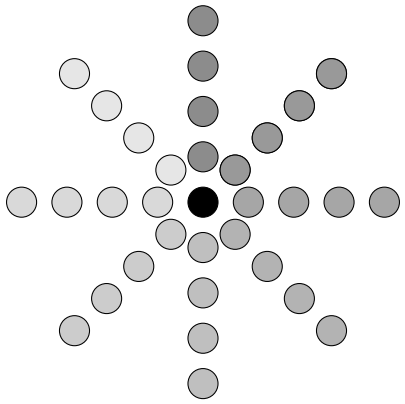


FIGURE 2.8 – Positionnement spatial des électrodes sur la langue de l'utilisateur. Les niveaux de gris indiquent la direction tandis que la localisation de l'électrode par rapport au centre de la langue représente la distance restante à parcourir [61].

2.2.2 Substitution auditive

Naturellement, les personnes voyantes peuvent assister les personnes malvoyantes lors de leurs déplacements à l'aide d'informations verbales. L'emploi de cette modalité informative est liée à la richesse à définir avec précision l'espace environnant et l'action à réaliser. Les méthodes d'assistance à la navigation basée sur une substitution auditive s'inscrivent dans la même continuité en apportant de manière autonome des informations pertinentes, telle la direction d'un chemin à suivre ou la position d'un obstacle, par le biais de messages vocaux définies à l'aide de phrases verbales ou de brèves indications (à gauche, à droite, etc) [67], [68]. D'autres méthodes, mettent à profit les caractéristiques intrinsèques de l'espace auditif pour encoder des informations à travers de brefs signaux

sonores afin d'éviter des délais d'émissions trop important pour un utilisateur malvoyant.

2.2.2.1 Perception auditive

L'ouïe est l'un sens de substitution naturel employé par les personnes malvoyantes pour comprendre leur environnement par la perception de bruits caractéristiques. Ces phénomènes auditifs sont des vibrations mécaniques qui se propagent sous forme d'ondes longitudinales à travers l'espace. La réception de ces vibrations par l'homme est effectuée par le tympan, une membrane fibreuse située à la jonction entre l'oreille externe et l'oreille moyenne. Lorsqu'un son atteint le tympan, il engendre des vibrations mécaniques qui font vibrer cette membrane. Les vibrations générées sont ensuite transmises dans l'oreille moyenne au moyen d'une chaîne d'osselets. Ces osselets amplifient les vibrations sonores et les transmettent ensuite à l'oreille interne. Au sein de l'oreille interne, se trouve la cochlée abritant des cellules sensorielles spéciales appelées cellules ciliées. Les vibrations mécaniques provoquent le mouvement des cils des cellules ciliées générant des signaux électriques. Ces signaux nerveux sont alors transmis par le nerf auditif vers le cerveau. L'ensemble de l'appareil auditif de l'oreille externe à interne est illustré à la figure 2.9.

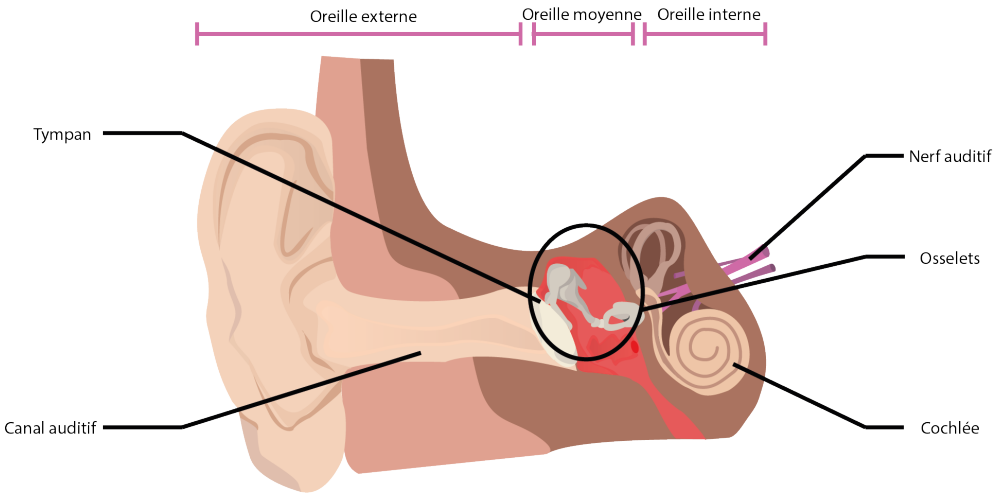


FIGURE 2.9 – Vue schématique de l'appareil auditif et de ces principaux éléments.

L'oreille humaine et le cerveau interprètent les caractéristiques fondamentales d'une onde acoustique pour la différencier d'une autre à partir de sa fréquence, de son intensité, de son timbre et de sa temporalité (définis ci-dessous). Néanmoins, la perception auditive humaine, tout comme celle des autres animaux, est limitée à certaines ondes mécaniques en raison des capacités intrinsèques de l'appareil auditif. Le champ auditif humain est compris communément entre 20 et 20 000 Hz, bien que l'oreille humaine soit plus sensible aux fréquences basses comprises entre 500 et 4 000 Hz. Un âge avancé

ou une exploitation plus poussée de notre perception auditive influe également sur nos capacités à percevoir certaines fréquences.

- **Fréquence :** Mesurée en *Hertz* ou *Hz*, elle dénote le nombre d'oscillations par seconde. Elle détermine la hauteur tonale d'un son : une faible fréquence évoque un son grave, tandis qu'une fréquence élevée correspond à un son aigu.
- **Intensité :** Exprimée en *décibel* ou *dB*, elle est corrélée à l'amplitude des oscillations sonores et traduit la puissance ou le volume sonore.
- **Timbre :** Ensemble des harmoniques d'un son le caractérisant.
- **Temporalité :** Aspect temporel d'un son, sa durée, son moment d'émission. La temporalité contribue à la perception d'une séquence sonore et au rythme musical.

En complément des indications de fréquence en Hertz, l'échelle de mels [69] peut être employée dans le cadre de la perception humaine pour décrire la fréquence d'un son. Tirant son nom du terme *mélodie*, cette échelle psychoacoustique est conçue pour mesurer la hauteur perçue des sons. Elle reflète la manière dont l'oreille humaine perçoit les variations de fréquence, offrant une résolution accrue dans les basses fréquences. Ainsi, elle a été conçue de manière que 1000 Hz correspondent à une valeur de 1 000 mels et qu'un rapport constant de la valeur en mels soit perçu par les auditeurs comme une variation constante de hauteur musicale. La conversion entre la fréquence en Hertz et l'échelle de mels est donnée par la formule 2.1.

$$\begin{aligned} mel &= 1127 \times ln(1 + \frac{freq}{700}) \\ freq &= 700 \times (e^{\frac{mel}{1127}} - 1) \end{aligned} \tag{2.1}$$

2.2.2.2 Localisation d'un signal sonore

L'oreille humaine permet de détecter des sons audibles et de discerner la direction d'où ils proviennent. Cette aptitude à localiser les sources sonores résulte d'un processus perceptif complexe, qui prend en compte à la fois les différences temporelles et les différences d'intensité sonore d'un même signal perçu par les deux oreilles [70]. La localisation sonore, en particulier selon l'axe horizontal, repose intrinsèquement sur deux mécanismes : la *différence temporelle interaurale (DTI)* (DTI ou *interaural time difference* -ITD), et la *différence d'intensité interaurale* (DII ou *interaural level difference* - ILD) [71].

La Disparité Temporelle Interaurale (DTI) fait référence à la différence de temps entre les arrivées d'un signal sonore aux deux oreilles. Cette différence, représentée

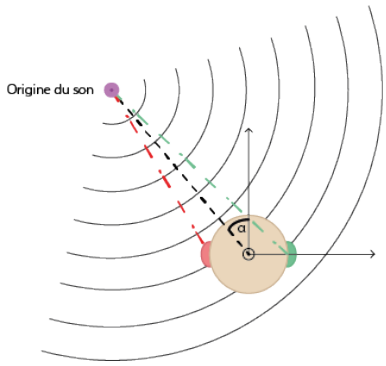


FIGURE 2.10 – Déplacement d'une onde acoustique dans un espace.

sur la figure 2.10, est due à une distance plus courte parcourue par le signal sonore vers l'oreille la plus proche de la source sonore. Elle peut également être dérivée d'une différence de phase entre les signaux sonores reçus, comme le montre l'équation 2.2, où φ_G et φ_D représentent les phases des signaux sonores respectivement pour l'oreille gauche et l'oreille droite. Une DTI de valeur nulle représente un bruit sonore d'origine frontale ou dorsale. La perception basée sur la DTI est plus efficace pour les sons dont la fréquence est inférieure à 1500 Hz. À partir de ce seuil, les retards de phase deviennent trop importants comparés à la longueur d'onde du signal.

$$DTI_p(\theta, f) = -\frac{\varphi_G - \varphi_D}{2\pi f} \tag{2.2}$$

L'DII, en revanche, désigne l'écart d'intensité sonore perçue par chaque oreille lorsque le son est émis d'une direction spécifique. L'atténuation du son par la tête et le corps en direction de l'oreille opposée entraîne une différence d'intensité entre les deux oreilles. Ce paramètre est exploité par le cerveau pour la localisation de la source sonore. Les fréquences élevées sont plus sensibles à l'DTI (> 2000 Hz) et sont ainsi complémentaires des DII malgré une faiblesse entre 1500 et 2000 Hz [72]. Une région nommée *cône de confusion* altère la précision de la localisation azimutale (figure 2.11). À l'intérieur de cette zone, dont l'axe coïncide avec la ligne interauriculaire, il en résulte une absence d'unicité sonore en termes de différences temporelles ou d'intensité [73]. Toutefois, des indices dynamiques tels que les mouvements de la tête viennent jouer un rôle crucial. Une inclinaison ou une rotation de la tête engendre des modifications dans l'intensité sonore et le délai temporel entre les oreilles, offrant ainsi à l'individu des informations concernant la véritable position spatiale de la source sonore [74].

La DII et la DTI servent exclusivement à déterminer la position azimutale de la source d'un son, tandis que la position verticale (élévation) est déterminée par sa composition

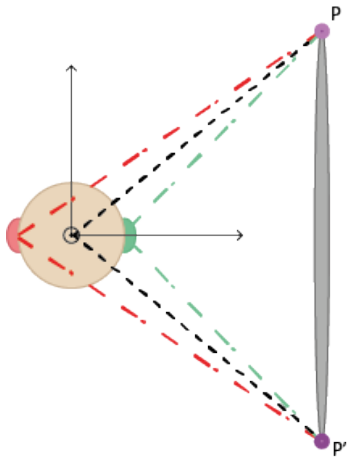


FIGURE 2.11 – Confusion de localisation sonore entre deux sons émis par deux émetteurs symétriques, perçus par l'oreille gauche (rouge) et droite (vert) d'une personne.

spectrale. En effet, au cours de son parcours dans l'espace environnant, un son rentre en interaction avec le corps humain en particulier, le torse, la tête et les pavillons des oreilles [75]. Les différentes zones de l'oreille et de la tête agissent comme des filtres acoustiques uniques, propres à chaque individu. Ces filtres modifient le spectre du son en fonction de sa provenance, soit en amplifiant certaines fréquences, soit en les réfléchissant ou en les absorbant. Ces modifications spectrales, qui sont transmises jusqu'au tympan, sont désignées par le terme HRTF (Head Related Transfer Function) permettent de localiser l'origine spatiale d'un son après un apprentissage continu par le cerveau tout au long de sa vie. Néanmoins, à l'instar, des DIIs et des DTIs, notre faculté de perception de l'élévation sonore présente des inégalités en fonction de la fréquence du signal. Notamment, la précision perceptive augmente avec des fréquences plus élevées [76], [77], malgré la présence d'indices acoustiques discernables dans les fréquences basses.

La perception de la distance entre une source sonore et un auditeur se révèle plus complexe, en particulier pour les personnes aveugles. L'absence concomitante d'information visuelle et sonore oblige l'auditeur à se fier uniquement à l'intensité du signal atténué en fonction de la distance parcourue. Malheureusement, l'intensité initiale d'un son varie substantiellement en fonction de sa source, rendant ainsi l'estimation de la distance ardue. Cependant, la localisation d'une source sonore dans un environnement, qu'il soit réverbérant ou libre, peut être compromise par la présence de bruits parasites. Ces interférences peuvent perturber le spectre audio de la source ciblée, en particulier si les caractéristiques acoustiques des interférences sont similaires à celles de la source [78]. La détermination de la localisation d'un ou plusieurs éléments dans un espace s'appuie sur un référentiel spatial. Dans le cas de l'espace auditif, mais également haptique, l'être humain s'appuie

sur un cadre **égocentrique** en partie lié à un manque d'information liant les éléments les uns aux autres [79]. Ce référentiel permet une représentation spatiale des objets relative par rapport à notre orientation sans connaissances sur leurs relations spatiales. À l'inverse, dans un référentiel allocentrique, utilisé notamment dans le contexte visuel, les objets sont localisés les uns par rapport aux autres (comme illustré à la figure 2.12).

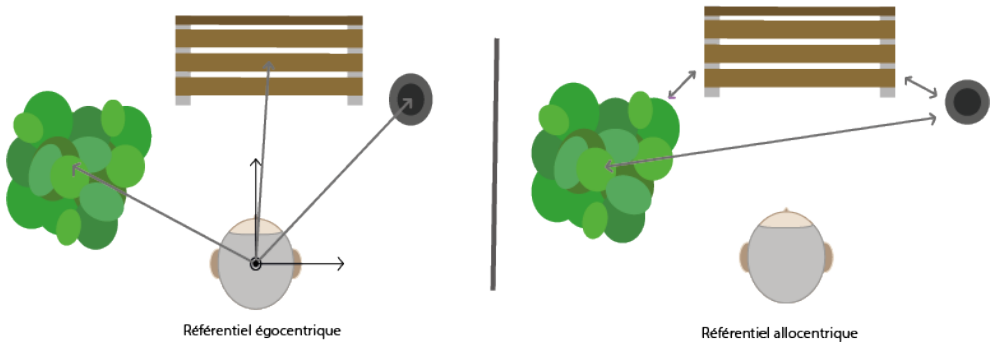


FIGURE 2.12 – Référentiel égocentrique versus allocentrique.

2.2.2.3 Espace visuo-spatial vers espace sonore

Spatialisation d'une image 2D par balayage :

Une des approches pionnières de transcription d'informations visuelles en stimulus sonores a été proposée en 1992 par Meijer [80]. Cette approche, nommée The vOICe, opère la transformation d'une image en niveaux de gris en séquences sonores, avec un encodage spécifique basé sur la temporalité, la fréquence et l'intensité du signal audio. Le processus de conversion, illustré dans la figure 2.13, effectue un balayage horizontal de chaque colonne de l'image, de manière séquentielle tel qu'une colonne correspond à un instant précis de l'émission sonore. Chaque pixel de la colonne est traduit en un signal sonore distinct, dont la fréquence oscille entre 500 et 5000 Hz. Une position verticale élevée du pixel se traduit par un son à fréquence élevée, tandis qu'une position inférieure engendre un son plus grave. De plus, l'intensité du signal audio est directement liée au niveau de gris des pixels. The vOiCe introduit le terme de "pixel sonore" définissant une relation entre la position d'un pixel au sein d'une image et un stimulus sonore unique, établissant ainsi un pont entre les dimensions visuo-spatiales et auditives. L'ensemble des "pixels sonores" d'une même colonne sont émis simultanément, créant ainsi une représentation sonore précise d'image dans un référentiel sonore. L'équation 2.3 décrit le signal sonore émis pour représenter le son associé à l'image à l'indice x à un instant t. L est la hauteur de l'image, f_y représente la fréquence relative à la position y, φ est

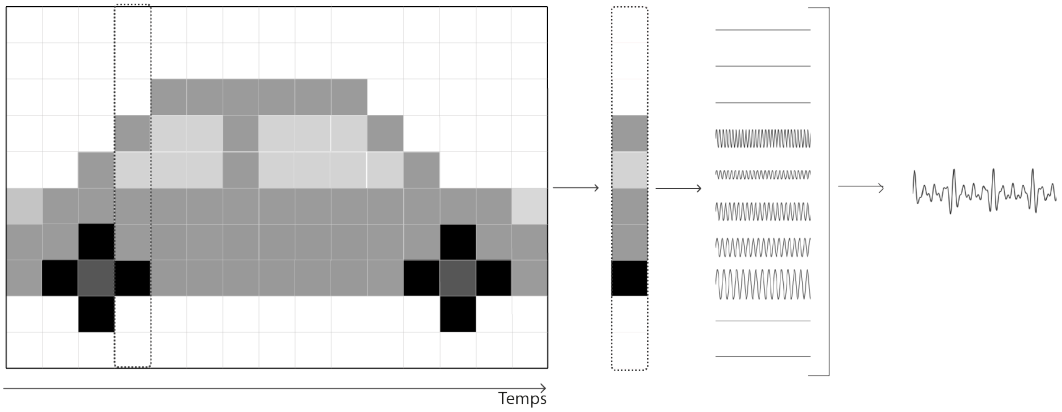


FIGURE 2.13 – The vOICe : Encodage d'une image 2D en signaux sonores à l'aide d'un encodage des pixels visuels en pixels sonore.

le déphasage initial du son, $g(x, y)$ indique la valeur du niveau de gris du pixel, et A est l'amplitude du signal sonore. Des adaptations ont été proposées au fil du temps suivant les évolutions technologiques [81].

$$s(x,t) = \sum_{y=0}^{L-1} A(g(x,y)) \sin(2\pi f_y t + \varphi) \tag{2.3}$$

Les techniques basées sur un balayage horizontal se sont avérées efficaces, après de longues périodes d'entraînement [82], pour détecter des formes simples ou complexes [83]. Elles sont également pertinentes dans la reconnaissance des orientations des lettres, au point que leur précision peut être mesurée via des tests ophtalmologiques comme le test de Snellen [81]. L'accès à des données supplémentaires sur l'environnement enrichit les informations, améliorant la précision de reconnaissance. En examinant l'importance de la couleur dans la perception visuelle des individus à vision normale, certaines méthodes incorporent cet aspect dans des systèmes sonores. Le dispositif *EyesMusic* [84] a prouvé une capacité significative à discriminer des formes élémentaires avec une précision de 91,5%. Cette approche transforme la couleur de chaque pixel en un son lié à un instrument musical spécifique, établissant ainsi un "dictionnaire" sonore. Après seulement 25 minutes de formation, des participants ont pu localiser avec précision et rapidité une position cible [85]. D'autres méthodes de substitution sensorielle traduisent la couleur d'un unique point d'intérêt de l'image [86] ou de plusieurs points en même temps [87]. Par exemple, *See ColOr* a permis à un individu malvoyant de suivre une ligne au sol sur 90 mètres [88]. Ces adaptations convertissent le code couleur (RVB ou HSV : Teinte, Saturation, Valeur) en une combinaison sonore unique basée sur la fréquence,

l'intensité et le timbre.

Spatialisation d'une image 2D via un signal monophonique :

L'analyse par balayage horizontal s'est révélée efficace dans la détection d'objets statiques. Cependant, cette méthode présente des contraintes pour décrire rapidement une information visuelle, notamment en raison d'importants délais pour parcourir l'ensemble de l'image. De ce fait, certaines méthodes de SSVA se concentrent sur la transmission d'une quantité limitée d'informations pertinentes via un signal sonore unique. *Prosthesis for substitution of vision by audition* ou *PSVA* [89] substitue le parcours horizontal en attribuant une fréquence unique, comprise entre 50 et 12526 Hz, pour chaque position verticale et horizontale de pixel. En outre, PSVA encode une information visuelle avec une densité de pixels sonore au centre de l'image supérieure à celle des extrémités, imitant la vision naturelle humaine. En effet, l'œil humain dispose d'une acuité visuelle maximale au centre, en particulier dans une région appelée la fovéa, qui est située au cœur de la macula. Cette zone est densément peuplée de cônes, des photorécepteurs spécialisés dans la vision des détails et des couleurs, ce qui lui confère une haute définition. La valeur de chaque pixel est transformée en une intensité sonore, similairement aux méthodes antérieures, mais est enrichie par une dimension stéréophonique où plus un pixel s'écarte horizontalement du centre de l'image, plus la différence d'amplitude perçue entre les deux oreilles est marquée. La stéréophonie est ainsi couramment utilisée pour localiser un élément horizontalement dans l'espace, alors que les variations de fréquence sont employées pour indiquer une position verticale. [90] utilise un décalage de phase entre les signaux sonores gauche et droite pour simuler une DTI artificielle jouant sur les capacités humaines de localisation horizontale à partir d'une différence temporelle. Des expérimentations avec des individus malvoyants montrent une aptitude à naviguer grâce à cette méthode pendant quatre jours [91]. D'autres méthodes utilisent des HRTFs artificielles reproduisant les perturbations ou transformations naturelles sonores afin de recréer un son spatial stéréophonique. Ces HRTFs synthétiques sont dérivées soit à partir de méthodes de synthèse numérique, soit par des mesures expérimentales employant des microphones intra-auriculaires, optimisés pour capturer fidèlement les variations acoustiques associées à différentes localisations spatiales autour d'entités réceptrices. La confrontation entre le spectre sonore source et le spectre capturé permet d'extrapoler les distorsions acoustiques survenues lors de la transmission sonore. Le processus d'enregistrement est produit dans une chambre anéchoïque (espace isolé de l'environnant extérieur avec des parois absorbantes des ondes mécaniques) afin d'éliminer

les réverbérations ou les bruits extérieurs pouvant parasiter les indices acoustiques reçus par la personne. L'exploitation de ces fonctions de transfert permet de générer des illusions auditives par la convolution d'un signal monophonique avec les HRTFs spécifiques aux oreilles gauche et droite décrit par l'équation 2.4, où *H* correspond à la paire de fonctions de transfert et *e* au signal monophonique. Ainsi, un son synthétique peut être généré pour obtenir une perception spatiale par l'utilisateur de cette source sonore virtuelle. Ce son généré à partir d'HRTF 2D permet de localiser cette source dans l'espace sonore [92], [93]. Par extension, une position dans l'espace visuel, par exemple celle d'un obstacle, peut être transposée dans cet espace sonore.

Cependant, l'utilisation de HRTF artificielle ou de personne tierce impacte grandement la précision de la localisation d'un stimulus sonore par rapport à des HRTFs reproduisant les transformations propres d'une personne. Dans une étude comparative mettant en lumière la précision de localisation [94] à partir des HRTFs génériques, celles d'une personne tierce et celles propres à l'individu testé, les précisions obtenues étaient respectivement de 44,12%, 55,07% et 83,56%. La localisation sur le plan horizontal reste constante avec une précision élevée dans l'ensemble des cas (98.67%, 98.67%, 100%). *See Differently* [95], exploite la stéréophonie basée sur une HRTF horizontale pour déterminer la position des objets environnants tout en conservant un indice fréquentiel pour décrire l'élévation. Un son monophonique distinct est utilisé pour décrire chaque type d'objet ou zone d'intérêt, permettant à l'utilisateur de contourner les obstacles ou de détecter la présence d'une route ou d'un chemin à proximité.

$$\begin{aligned} s_G(\theta, \phi, t) &= H_G(\theta, \phi) * e_0(t) \\ s_D(\theta, \phi, t) &= H_D(\theta, \phi) * e_0(t) \end{aligned} \tag{2.4}$$

Spatialisation sonore 3D

La distance séparant un point d'intérêt de l'espace et l'utilisateur, bien que difficilement perceptible naturellement, est un élément important pour appréhender son environnement avec une localisation de la proximité des objets. L'encodage de la distance s'appuie principalement sur le mécanisme naturel humain pour estimer la distance d'un son, c'est-à-dire son atténuation sonore. Les données visuelles 3D, issue de la carte de profondeur ou d'un nuage de points, module le signal sonore résultant d'une spatialisation 2D pour obtenir une 3ème dimension d'une manière analogue aux mécanismes utilisés pour représenter les gradations de gris lors de la conversion d'une image bidimensionnelle [92], [96]-[98]. De plus, la normalisation de l'intensité sonore des sons artificiels ne subit pas

les disparités naturelles de puissance sonore des éléments de l'espace, compliquant la perception de distance par rapport à son origine. Les indices acoustiques contenus dans la plage spectrale d'un son fournissent à l'homme des informations supplémentaires relatives à la distance par rapport à la source sonore. Ces indices peuvent ainsi être reproduits avec des HRTFs 3D, enregistrées à de multiples distances, pour simuler une émission plus ou moins éloignées [99].

2.3 Vision artificielle dans les systèmes de substitution sensorielle

Les capacités humaines de perception de stimulus sonore ou haptique, telles qu'évoquées antérieurement, et de localisation spatiale d'une ou de multiples sources sensorielles de manière simultanée, ont ouvert la voie à l'augmentation de l'information sensorielle naturelle grâce à des sources synthétiques en vue d'améliorer le quotidien des personnes malvoyantes en particulier dans leurs déplacements. L'absence de donnée visuo-spatiale, fondamentales pour les personnes voyantes, prive les individus non-voyants d'une appréhension complète et structurée de leur environnement. L'apport additionnel d'informations visuelles capturées à partir de caméra monoscopique, stéréoscopique ou par imagerie LIDAR peuvent permettre de compenser en partie cette déficiente en renseignant la personne non-voyante sur des potentialités de danger, comme la présence d'un escalier [100], ou même de déterminer un chemin à emprunter dans un couloir pour esquiver les obstacles, garantissant ainsi une navigation plus fluide et sécurisée. Ces informations peuvent être extraites des données visuelles et soumises à une analyse via des techniques de vision par ordinateur, en particulier des méthodes d'apprentissage automatique, afin de les rendre intelligibles par une réduction de leurs complexités. Les méthodes de vision artificielles ont été développées à partir des années 1960 pour substituer l'être humain dans l'analyse d'informations visuelles ou d'accéder à des informations non visibles naturellement. L'une des premières applications de ces méthodes fut dans le domaine spatial pour restaurer et améliorer le rendu d'images lunaires capturées, avant de s'étendre à la médecine pour l'imagerie médicale, et même au grand public, révolutionnant la qualité des photos numériques. L'émergence des méthodes par apprentissage, une branche des techniques d'intelligence artificielle, ont permis de traiter de manière automatique l'information visuelle et donc d'accéder à des informations inaccessibles jusqu'à alors, telles que les informations utiles à la navigation d'une personne visuellement déficiente. Dans la continuité de ce sous-chapitre, nous présenterons d'abord une vue d'ensemble des mé-

thodes d'apprentissage automatique et d'apprentissage profond dédiées à l'identification d'éléments clés dans une scène visuelle. Ensuite, nous explorerons leurs applications au sein des systèmes d'assistance visuelle.

2.3.1 Méthode d'apprentissage automatique (Machine learning)

Les méthodes d'apprentissage automatique ou *Machine Learning (ML)*, constituent un sous-domaine spécialisé de l'intelligence artificielle (IA) qui s'appuie sur des modèles statistiques pour résoudre divers problèmes. Ces méthodes se distinguent par leurs capacités à exploiter des données d'entraînement lors d'une phase d'apprentissage pour en extraire des modèles prédictifs ou des règles, plutôt que de s'appuyer sur une programmation définie lors de sa conception. Un algorithme d'apprentissage est employé pour définir automatiquement une modélisation du problème à partir de l'optimisation d'une fonction objective, souvent définie en termes de coût ou d'erreur, visant à affiner la pertinence du modèle à mesure de la présentation de nouvelle donnée d'entrée. Selon la nature du problème et les données disponibles, plusieurs méthodologies d'apprentissage peuvent être déployées : apprentissage supervisé, non supervisé ou encore apprentissage par renforcement.

- **Apprentissage supervisé :** Approche d'apprentissage d'un modèle algorithmique f à partir de la connaissance d'un ensemble de données $\mathcal{D}$ où pour chaque valeur d'entrée x_i, une valeur de sortie y_i, nommée étiquette, est associée. L'objectif est de définir et d'affiner la relation de f telle que la prédiction de $f(x_i)$, nommée $\hat{y}$, soit proche de la valeur y_i (figure 2.14).

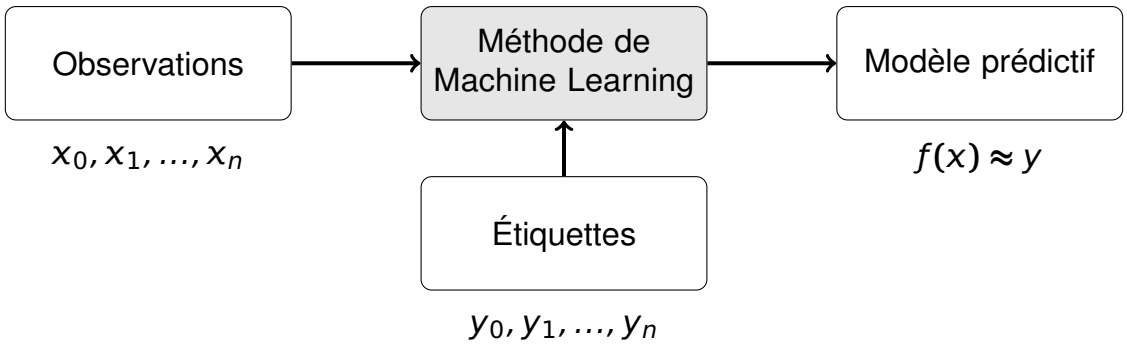


FIGURE 2.14 – Processus d'apprentissage supervisé.

Les méthodes d'apprentissage supervisé sont impliquées dans des problèmes de régression et de classification. Dans le cas des problèmes de régression, les étiquettes

ainsi que les valeurs prédites sont définies sur l'ensemble continu $\mathbb{R}$. À l'opposé, pour les tâches de classification, elles sont déterminées par un ensemble discret d'entiers $\mathcal{C}$ symbolisant les différentes classes possibles. Lorsque $\mathcal{C} = 2$, on est en présence d'une classification binaire. Pour $\mathcal{C} > 2$, la classification est multi-classe.

- **Apprentissage non supervisé :** Apprentissage à partir d'un ensemble de données $\mathcal{D} = x_0, x_1, ..., x_n$ non étiqueté afin de définir de manière autonome des structures ou des relations communes entre individus. Cet apprentissage est utilisé dans des méthodes de partitionnement *(clustering)* associant des éléments partageant des similarités ou de réduction de dimension représentant les données dans un espace de dimension plus faible que celui d'origine (figure 2.15).

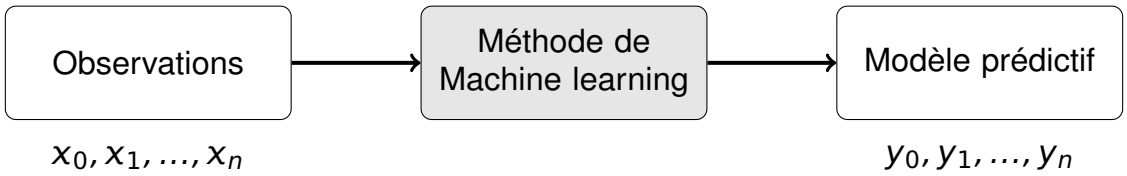


FIGURE 2.15 – Processus d'apprentissage non supervisé.

- **Apprentissage semi-supervisé :** Méthode hybride évoluée entre les méthodes avec ou non-supervision basée sur un ensemble de données non globalement étiquetée. Un jeu de données non entièrement étiqueté peut-être dû à un accès important à de nouvelles observations sans possibilité de toutes les annotées, soit en raison d'une impossibilité d'accéder à l'information requise, soit en raison de la nature chronophage du processus d'annotation. Un apprentissage strictement supervisé peut introduire des biais, notamment en raison de potentielles erreurs d'étiquetage. En revanche, l'exploitation d'observations non annotées facilite la découverte automatique de structures ou de motifs communs au sein des données. La combinaison des deux approches, supervisée et non supervisée, tend à améliorer la capacité de généralisation du modèle face à de nouvelles observations.
- **Apprentissage par renforcement :** Méthode d'apprentissage où un agent évolue en interagissant dynamiquement avec son environnement. Le comportement de l'agent est progressivement affiné à travers un processus itératif afin d'optimiser une fonction objective. Le processus itératif s'appuie sur un mécanisme de retour d'expérience, basé sur des métriques de récompense et de pénalité, qui guide l'évolution de l'apprentissage. Les applications typiques comprennent les jeux, la navigation robotique et l'optimisation de processus complexes.

L'apprentissage automatique et les données sont le socle d'un fonctionnement

approprié de la machine. Des données inadaptées ou biaisées impactent la précision de modélisation du phénomène. De plus, le sous-apprentissage, ou *under-fitting*, se manifeste par une mauvaise adaptation du modèle aux données, tandis que le sur-apprentissage (*over-fitting*) résulte d'un apprentissage trop spécifique aux données d'entraînement, rendant le modèle incapable de bien généraliser sur de nouvelles données. L'observation de ces phénomènes est réalisée par une scission de l'ensemble des observations en deux sous-ensembles : un pour l'entraînement et l'évaluation du modèle nommé *jeu de donnée d'entraînement*, et un autre nommé *jeu de donnée de validation* pour évaluer ses performances à partir de métriques spécifiques au problème (détaillé à l'annexe 7.3). Des performances faibles sur l'ensemble des groupes de données représentent un modèle mal adapté au problème pouvant résulter d'un sous-apprentissage, tandis qu'un écart significatif de performance entre les données d'entraînement et les données de validation indiquent un problème de généralisation.

Les méthodes de ML adressent de multiples applications ou des problèmes comme la classification et le clustering. Le clustering est une décomposition d'un ensemble d'individus en sous-groupes homogènes partageant des structures ou des traits caractéristiques communs définis à partir de critères de similitude tels qu'un calcul de distance entre deux éléments ou bien une métrique de similarité. Une tâche de classification associe à une observation une étiquette à partir d'un ensemble de possibilités, nommée classe ou catégorie préalablement définie, représentant la nature d'un objet, une propriété caractéristique des données. Dans le cas d'une classification binaire, le système se limite à deux issues possibles. Cette dichotomie s'apparente à une question fermée donc la réponse est vrai/faux. En revanche, la classification multi-classes est analogue à une question ouverte telle qu'une question à choix multiples (QCM) où une unique réponse est requise par question, par exemple, une interrogation du type : *"Identifier le genre de véhicule représenté dans cette image : est-ce une voiture, un bus ou bien aucun véhicule n'est présent ?"*. D'autres méthodes, nommées méthodes de classification multi-étiquettes, offrent une possibilité de réponses multiples.

Dans le cadre de méthodes de clustering ou de classification d'éléments par ML, une chaîne de traitement est effectuée en amont pour extraire des informations brutes des images des exploitables permettant de quantifier ou de mesurer une similitude avec les données apprises lors de l'apprentissage. La figure 2.16, représente le processus global, de l'image brute à la prédiction de sortie. L'opération initiale consiste à extraire des régions d'intérêt, ou ROI (*Region of Interest*), pour identifier des segments spécifiques de l'image et distinguer divers éléments ainsi que leurs positions relatives au sein de celle-ci à partir d'un partitionnement régulier de l'image en fenêtre. Les informations importantes

des ROI sont ensuite extraites à l'aide de techniques d'extraction de caractéristiques afin de les rendre exploitable sous forme de descripteurs (vecteur ou ensemble de données représentant une ou plusieurs caractéristiques différentes). Une méthode de sélection des caractéristiques les plus discriminantes peut être appliquée pour diminuer la quantité d'informations encapsulées dans les descripteurs. Enfin, ces informations alimentent une méthode de ML afin de produire une prédiction en se basant sur les connaissances acquises lors de la phase d'apprentissage.

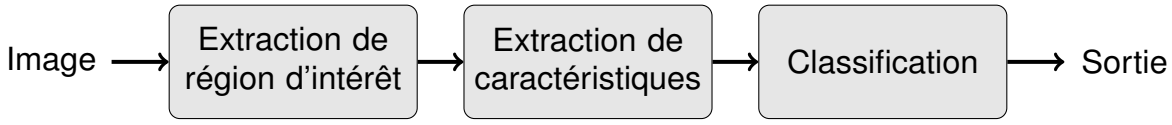


FIGURE 2.16 – Étape d'une classification par ML.

2.3.1.1 Extraction de caractéristiques

L'extraction de caractéristiques transforme des données d'images brutes et les informations exploitables pour permettre aux méthodes de ML d'associer des éléments similaires à partir de leurs caractéristiques partagées comme sa forme géométrique, sa couleur ou sa texture [101]. Une couleur est définie par le contenu spectral de l'image communément défini par trois canaux chromatiques (Rouge, Vert, Bleu $\sim$ RVB) tandis que la texture d'une image est une variation de l'intensité autour d'un pixel ou un petit motif structuré élémentaire et répétitif caractérisant une surface d'un objet. L'ensemble de ces caractéristiques peut être défini en fonction de leur perspective, qu'elles soient locales ou globales. Une méthode locale définit un ensemble de caractéristiques à partir de multiples points de vue de l'image en s'intéressant uniquement aux informations les plus proches afin d'augmenter la robustesse aux occlusions partielles de l'élément dans la définition des propriétés. À l'opposé, une méthode globale généralise l'ensemble de l'image à partir d'un unique vecteur de caractéristiques limitant l'impacte de faibles variations lumineuses ou du mouvement.

De nombreuses méthodes globales existent pour définir des caractéristiques de l'image au sein de vecteurs de caractéristiques. Les informations statistiques de couleurs génèrent des descripteurs sur la distribution de fréquence des différentes teintes dans une image à partir d'histogrammes renseignant sur la moyenne, la variance et l'asymétrie des distributions colorimétriques (Color moments). La nature périodique d'une texture est caractérisée par des méthodes statistiques ou fréquentielles. La matrice de co-occurrence des niveaux de gris évalue la fréquence à laquelle des paires spécifiques de pixels de luminosité ou de couleur se produisent à une distance et orientation données. D'autre part,

l'analyse fréquentielle, comme la transformée de Fourier, décompose l'image en composantes fréquences pour détecter les motifs répétitifs, comme ceux inhérents aux textures. Les moments invariants sont largement employés pour identifier la forme géométrique d'un objet dans une image binarisée due à leurs caractéristiques de robustesse aux variations de translation, la rotation et l'échelle [102].

Les méthodes locales se basent également sur les propriétés colorimétriques ou géométriques mais sur une échelle faible ou basse (*Low-level image*) c'est-à-dire non plus sur l'aspect global de l'élément, mais sur une zone réduite de l'image. Des informations caractéristiques de la texture sont ainsi analysées par des techniques basées sur la reconnaissance de motifs binaires locaux (Local Binary Pattern ∼ LBP). Des extensions de LBP ont été proposées dans la continuité pour réduire la dimensionnalité des descripteurs (Motif Binaire uniform ∼ Uniform Binary Pattern) ou pour augmenter la robustesse ou son efficacité en utilisant non plus une analyse pixel par pixel, mais par groupe de pixels (Three-Patch LBP). D'autres méthodes telles que les filtres de Canny ou de Sobel définissent des vecteurs de caractéristiques sur la présence des frontières, mais les méthodes d'assistances visuelles implémentent principalement les méthodes par histogramme orienté de gradient (HOG) [103] ou par point-clé tel que SIFT [104] (Scale-Invariant Feature Transform) et ses extensions comme SURF (Speeded-Up Robust Features), FAST (Features from Accelerated Segment Test) ou ORB (Oriented FAST and Rotated BRIEF) encapsulant leurs informations sous forme de descripteurs.

2.3.1.2 Classification

Les descripteurs ou vecteurs de caractéristiques, définis par le biais d'algorithmes d'extraction et de sélection, encapsulent des informations locales ou globales relatives aux propriétés de l'image. La phase de classification vise à convertir ces informations désormais utilisables en une prédiction y parmi un ensemble de classes possibles. Une approche simple consiste à évaluer la similitude entre le vecteur inconnu x et un ensemble de vecteurs caractéristiques étiquetés $\mathcal{X}$ par une métrique de distance, définie en annexe 7.1. La classe du vecteur de caractéristiques étiqueté le plus proche est par extension le plus similaire est alors attribuée au vecteur inconnu. La méthode des K-plus proches voisins (KNN) étend ce processus non seulement sur la base du voisin le plus proche, mais par rapport l'étiquette majoritaire parmi les K échantillons les plus proches, atténuant ainsi l'influence d'une donnée aberrante liée à du bruit.

Des méthodes de ML statistiques définissent lors de l'apprentissage des séparateurs linéaire ou non pour diviser l'ensemble de vecteur $\mathcal{X}$ annoté en sous groupe partageant une classe commune. Un séparateur linéaire, donc le classifieur linéaire en est l'exemple

le plus simple, permet de segmenter l'espace à partir de segments, de droites, de plans ou d'hyperplans selon la dimensionnalité. Le nombre de séparateurs linéaire peut-être différent suivant le nombre de classes à segmenter. Les séparateurs à vaste marge (SVM - *Support vector machine*) reposent sur la recherche du séparateur linéaire avec la marge la plus importante. Cette marge, optimisée pour diminuer l'erreur de bruit des données, est délimitée par des vecteurs spécifiques appelés "vecteurs de support", comme le montre la figure 2.17. Néanmoins, dans de nombreux scénarios pratiques, la non-linéarité des données, due à des bruits ou à des chevauchements, requiert des adaptations. Une marge souple est alors introduite dans le SVM pour tolérer certaines erreurs de classification, et lorsque la séparation linéaire est tout simplement impossible, des fonctions de noyau ou *kernel* (Gaussien, Polynomial, Fonction de base radiale, etc) projettent les données dans un espace de dimension plus élevé afin de rendre la séparation linéaire réalisable. Les espaces non linéairement séparables peuvent également être classifiés à l'aide d'arbre de décisions, donc la prédiction est réalisée par une suite de décisions basée sur les valeurs des caractéristiques.

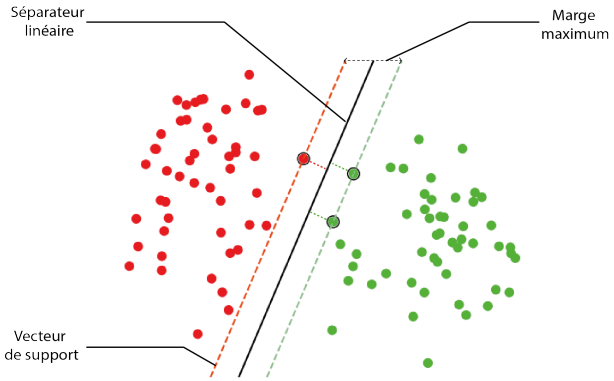


FIGURE 2.17 – Séparation linéaire par SVM.

Les arbres de décision (*decision tree*) sont utilisés pour résoudre des problèmes où les données ne peuvent pas être divisées linéairement à partir d'une série de décisions hiérarchiques basées sur les attributs des données. Le terme "arbre" est employé en raison de la ressemblance structurelle avec les arbres biologiques, incluant des nœuds et des branches. Dans ce contexte, chaque nœud représente un test sur une caractéristique spécifique, et les branches issues de ces nœuds mènent soit à un autre nœud pour une évaluation supplémentaire, soit à une feuille, qui est le résultat final ou la décision prise par l'arbre. La construction de l'arbre est effectuée généralement à partir de la racine en utilisant l'ensemble de données complet. À chaque nœud, un algorithme de construction, tel que l'algorithme Classification And Regression Trees (CART) [105], sélectionne

l'attribut qui offre la meilleure séparation possible des données en fonction d'un critère d'homogénéité. Une fois cet attribut choisi, les données sont divisées en sous-ensembles en fonction des résultats du test par cet attribut. Ces sous-ensembles servent ensuite à créer les nœuds enfants du nœud actuel jusqu'à atteindre une profondeur d'arbre limite ou avoir un nombre minimal d'échantillons dans un nœud. Une manière d'augmenter la robustesse et la fiabilité de cette prise de décision consiste à évoluer d'un simple arbre de décision à une forêt d'arbres, connue sous le nom de forêts aléatoires (*random forest*) ou forêts décisionnelles. Ces forêts fusionnent les prédictions de multiples arbres de décision autonomes, pour aboutir à une prédiction globale, qui est souvent la conséquence d'un consensus ou du vote majoritaire de tous les arbres inclus.

2.3.2 Apprentissage profond

Les techniques d'apprentissage profond (*Deep-Learning*) ont vu leurs applications augmenter considérablement dans le domaine de la vision par ordinateur, principalement poussée par des progrès surpassant les méthodes traditionnelles. Le "deep learning" constituent une spécialisation de l'intelligence artificielle et du machine learning, s'appuyant essentiellement sur des architectures de réseaux de neurones artificiels (ANN - Artificial Neural Network). Le concept de neurone artificiel trouve ses racines dans le "neurone formel" défini par McCulloch et Pitts [106] et illustré à la figure 2.18. Ce concept fut enrichi par Rosenblatt avec une méthode d'apprentissage automatique pour former un *perceptron* [107]. Sa structure est une abstraction simplifiée d'un neurone biologique, conceptualisant le flux d'information à travers des signaux binaires, analogues à l'absence ou la présence de signaux nerveux. Le neurone formel, dans son fonctionnement, effectue des opérations sur ses entrées pour générer une sortie unique. Il se compose de cinq composants majeurs, à savoir :

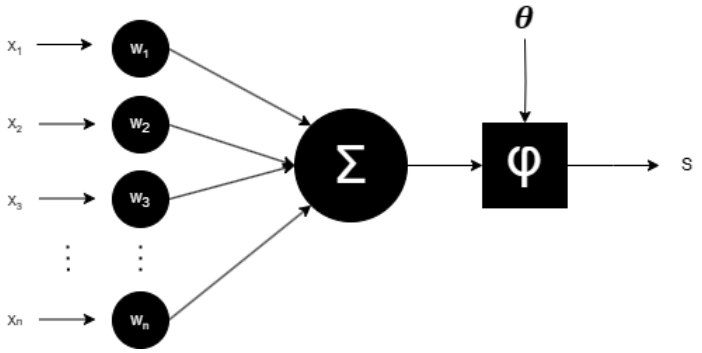


FIGURE 2.18 – Représentation graphique d'un neurone formel.

- **Couche d'entrée :** Accueille les données d'entrée brutes, à l'instar des dendrites des neurones biologiques. L'information perçue à chaque neurone est pondérée par un poids synaptique.
- **Bias :** Introduit une constante au résultat de la fonction d'activation pour ajuster sa dynamique.
- **Fonction de transfert :** Calcule la somme pondérée des entrées plus un biais avec les poids synaptiques associés.

$$f(x) = \sum_{i=1}^{n}(x_n w_n) + b = W^T.X + b \tag{2.5}$$

- **Fonction d'Activation :** Fonction mathématique non linéaire qui détermine l'activation du neurone. Pitts & McCulloch ont introduit une fonction de seuil, mais d'autres fonctions, telles que la sigmoïde pour normaliser un résultat, peuvent être employées.
- **Neurone de sortie :** Représente la valeur finale émise par le réseau. Dans le contexte du neurone binaire, elle est similaire à l'émission ou à l'absence d'un signal électrique par un axone.

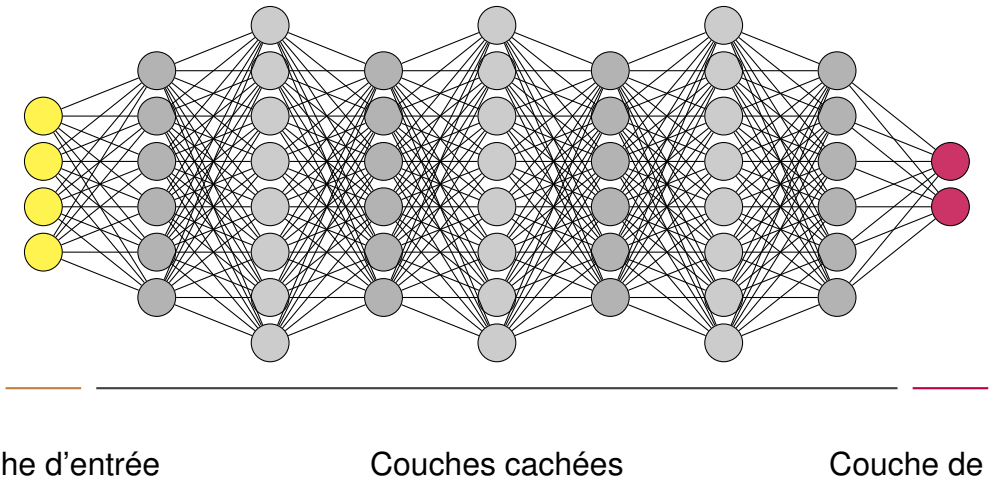


FIGURE 2.19 – Réseau de neurone multicouches.

Un neurone formel isolé présente des limitations pour traiter des problèmes d'une grande complexité, multidimensionnelles ou non-linéaires, principalement due à un manque de paramètre et d'opération. Néanmoins, cette approche élémentaire est le fondement de réseaux neuronaux plus avancés composés de multiples neurone formel ordonnées en parallèle et couches successives. Cette disposition hiérarchisée, bien qu'abstraite, s'inspire

des connexions cérébrales où les neurones biologiques, étroitement interconnectés, orchestrent une transformation et une transmission séquentielle et parallèle de l'information. Un modèle neuronal à plusieurs couches, couramment utilisé en apprentissage profond, se compose typiquement d'une couche d'entrée, de diverses couches cachées et d'une couche de sortie (figure 2.19). L'information est continuellement transformée au fil de son passage à travers les couches. La force des connexions neuronales est déterminée par les poids synaptiques rendant certains neurones plus prépondérants dans le processus décisionnels. Cependant, une mauvaise pondération altère les capacités de traitement des informations en donnant un poids important à des éléments non pertinents. Des méthodes d'apprentissages supervisés sont ainsi utilisées pour optimiser les paramètres de modèles tels que les poids synaptiques W et les biais b. Le processus d'apprentissage vise à minimiser une fonction de coût $\mathcal{J}$ définie à chaque itération par l'écart entre les prédictions $\hat{y}$ et les valeurs réelles y associées à un ensemble d'entrées x. L'équation 2.6 de la fonction de coût intègrent une fonction de quantification de l'écart nommée fonction d'erreur $\mathcal{L}$ (Défini plus en détail en annexe 7.2) et une fonction de régulation g visant à assurer la généralisation du modèle. Donc son poids sur le coût total est modulée par le coefficient λ.

$$\mathcal{J}(W, b) = \mathcal{L}(\hat{y}(x, W, b), y) + \lambda g(W, b) \tag{2.6}$$

Les paramètres des modèles, au début de l'apprentissage, sont initialisés par une distribution aléatoire conformément à des méthodes spécifiquement élaborées pour garantir une convergence efficace et stable lors des phases d'entraînement. Puis, à chaque itération de l'apprentissage, ces paramètres sont ajustés de manière plus ou moins significative par une méthode de rétropropagation du gradient en fonction de la valeur coût. Les algorithmes de rétropropagation du gradient comme la descente de gradient stochastique ou des variantes plus sophistiquées, telles qu'Adam et RMSprop sont des méthodes de modification des paramètres du réseau où l'information de coût est diffusée rétroactivement à travers le réseau depuis la couche de sortie jusqu'à la couche d'entrée.

Comme évoquée précédemment, l'information en entrée d'un réseau est véhiculée de manière linéaire de couche en couche jusqu'à atteindre celle de sortie. Ce type de rseau est nommée *Réseaux de neurones à propagation avant* (*feedforward neural network* ∼ FNN). Cette transmission linéaire des données à travers les couches assure que chaque entrée est traitée de manière indépendante, sans références à des entrées antérieures ou subséquentes. D'autres méthodes modifient la transmission linéaire des données par une transmission bidirectionnelle où les données antérieures influencent le comportement courant du réseau. Ces réseaux nommés *Réseaux neuronaux récurrents* ∼

RNN maintiennent une "mémoire" des entrées précédentes dans leurs séquences à partir de boucles de récurrence qui renvoient l'information à l'entrée du même neurone. Cette mémoire offre un état interne reflétant l'historique des entrées traitées. Cette récurrence temporelle convient au traitement de données séquentielles, où des interdépendances entre les éléments existent. Ces deux structures fondamentales sont le socle de nombreux types de réseaux neuronaux profonds tel que :

- **Perceptron multicouche ∼ MLP :** est un réseau feedforward classique où chaque neurone d'une couche est connecté à tous les neurones de la couche suivante. Cette particularité est qualifiée de *fully-connected* ou *dense*. Cette architecture permet au MLP d'être particulièrement adapté pour traiter des situations sur lesquelles les données ne sont pas linéairement séparables, liée à la présence de couches cachées combinées à des fonctions d'activation non linéaires.
- **Réseau neuronal convolutionels ∼ CNN :** est un réseau feedforward classique largement répandu dans le domaine de la vision artificielle. Les réseaux CNN substituent les couches denses par des couches de convolution. Des opérations de convolution successives sont effectuées pour déterminer des motifs locaux pertinents au sein des données. Les poids synaptiques des neurones sont remplacés par des paramètres de noyau de convolutions.
- **Réseau récurrent à mémoire court et long terme ∼ LSTM :** est une variante des RNN composée de cellule LSTM, conçue pour répondre aux problèmes rencontrés par ceux-ci, de disparition du gradient lors de l'apprentissage. Les cellules LSTM ont pour caractéristique de maintenir un état aussi longtemps que nécessaire à partir d'une 2ème entrée activant ou non la fonction d'oubli de la cellule. Les méthodes LSTM sont fréquemment utilisées pour traiter des séquences d'informations interdépendantes, comme pour les textes ou les vidéos.

Dans la suite de ce manuscrit, nous nous concentrerons sur les approches basées sur les réseaux neuronaux convolutifs, largement utilisées dans le domaine de la vision par ordinateur et, par conséquent, dans les systèmes d'assistance visuelle pour les personnes malvoyantes.

2.3.2.1 Réseau de neurone convolutionnels

Les réseaux de neurones convolutionnels, apparus dans les années 1980, sont une spécialisation des réseaux neuronaux profonds principalement conçus pour la vision artificielle et l'analyse d'informations visuelles afin de palier des limitations inhérentes aux modèles neuronaux. Les couches denses des modèles neuronaux, tels que les perceptrons

multicouches, négligent l'importance de l'arrangement spatial des pixels d'une image lors de leur traitement, transformant l'information matricielle de l'image en vecteur. En outre, les données visuelles entraînent une quantité importante de paramètres dès la couche d'entrée, une situation aggravée par la superposition de nombreuses couches denses. Ce phénomène prolonge et complique la phase d'apprentissage à cause de la croissance rapide du total des paramètres. Par conséquent, la résolution de l'image d'entrée doit parfois être réduite pour restreindre l'explosion paramétrique pouvant rendre certaines informations cruciales non discernables. En revanche, les couches de convolution et de poolings des CNNs réduisent considérablement le nombre de paramètres tout en préservant la structure spatiale des pixels de l'image.

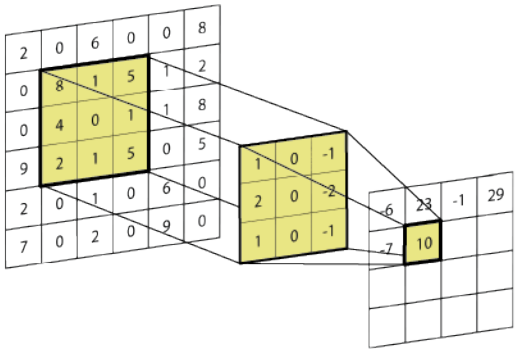


FIGURE 2.20 – Principe de la convolution.

Une couche de convolution applique un opérateur de convolution à une entrée matricielle avec un filtre, nommée noyau de convolution, de taille inférieure pour couvrir un champ réceptif spécifique de l'image. L'opérateur de convolution balaie le champ réceptif de l'image et réalise des multiplications élément par élément entre la section d'image en cours et les valeurs du noyau avant d'être agrégé pour obtenir une valeur unique. Cette opération est reproduite sur l'ensemble de l'image, en glissant le noyau de convolution à travers différents champs réceptifs afin de créer une carte de caractéristique comme illustrée à la figure 2.20. La carte des caractéristiques obtenue à partir d'un unique noyau de convolution englobe des motifs locaux de nature uniforme avec un nombre réduit de paramètre. L'application simultanée de plusieurs noyaux dans une seule couche de convolution extraient différents attributs et nuances de l'entrée tels que les contours et la texture. Par ailleurs, l'opération de pooling ou de regroupement vise à réduire les dimensions des cartes de caractéristiques par des opérateurs locaux ou globaux. Cette réduction présente deux bénéfices majeurs, une diminution du nombre de paramètres nécessaires et l'ajout d'une invariance aux translations.

Les méthodes par CNN se distinguent des techniques par Machine Learning en termes d'architecture. Contrairement aux méthodes par Machine Learning qui s'appuient sur une extraction manuelle préalable des caractéristiques via des descripteurs spécifiques tels que SIFT ou HoG, les CNN opèrent dans une perspective end-to-end. Cette perspective englobe à la fois le processus d'extraction des caractéristiques et celui de la classification au sein d'une unique approche, comme illustré dans la figure 2.21. L'extraction automatique des caractéristiques est réalisée par une série de couches, majoritairement convolutionnelles, qui permettent d'obtenir des représentations des données de plus en plus abstraites et pertinentes. Cette séquence de couches est communément appelée *backbone*. Après cette étape d'extraction, une opération de mise à plat, ou "flattening", est appliquée, convertissant ainsi les matrices multidimensionnelles des cartes de caractéristiques en un vecteur linéaire. Ce vecteur est ensuite propagé à travers des couches entièrement connectées, souvent associées à des fonctions d'activation non-linéaires, pour prédire une étiquette de classe. Les réseaux de CNN, outre une application de classification, permettent des applications plus complexes telles que la localisation d'un ensemble d'éléments pertinents préalablement définis lors de l'apprentissage ou une classification de l'image par pixel nommée segmentation sémantique. Des méthodes évoluées entremêlant les deux aspects nommés détection et segmentation d'instance permettre de distinguer et d'identifier plusieurs occurrences du même objet séparément.

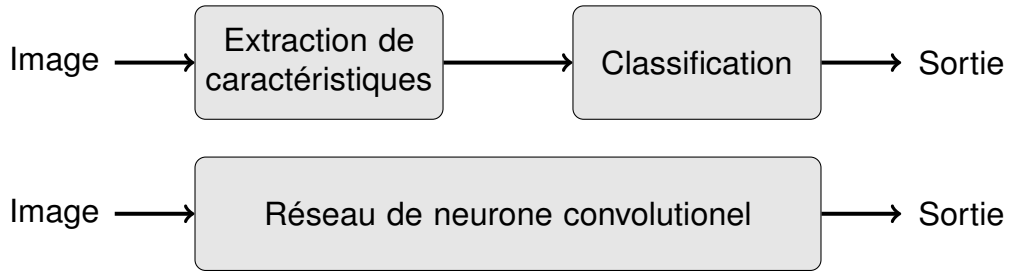


FIGURE 2.21 – Différence structurelle entre les méthodes par ML (en haut) et par CNN (en bas).

2.3.2.2 Détection d'objets

La détection d'objets est une discipline qui conjugue deux composantes majeures de la vision par ordinateur : la localisation et la classification. La localisation se charge de définir la position d'un ou plusieurs éléments dans une image, tandis que la classification associe à chaque élément détecté une catégorie ou une classe spécifique. Ces informations sont typiquement représentées par des vecteurs contenant les coordonnées d'une boîte englobante (en 2D ou 3D) et un identifiant correspondant à la classe de l'objet, déterminé à partir d'un ensemble de classes définies lors de la phase d'entraînement. La

figure 2.22 illustre une scène extérieure où différents éléments sont localisés et catégorisés par des couleurs distinctes selon le type d'objet. Historiquement, les premières méthodes pour définir la position d'un objet dans une image utilisaient une approche de fenêtre glissante pour analyser morceau par morceau une image avec une méthode de classification par ML puis CNN [108]. Cependant, en raison de la multitude de fenêtres à considérer, ces méthodes se révélaient particulièrement coûteuses en termes de calcul, limitant leur applicabilité sur d'importants volumes de données. Depuis, deux grandes catégories de méthodes de détection d'objets automatique se sont distinguées. La première s'appuie sur la proposition d'objets, où des zones saillantes de l'image sont extraites comme candidats potentiels à la détection, puis une seconde, plus récente et généralement plus rapide, reposant sur des méthodes de régression qui estiment directement la position et la catégorie des objets dans l'image.

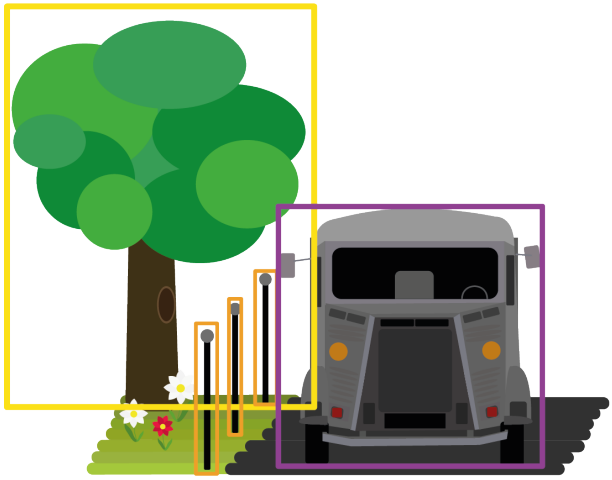


FIGURE 2.22 – Détection d'obstacle par boites englobantes.

Les techniques de détection d'objets basées sur la proposition de régions d'intérêt (ROI) se distinguent des approches traditionnelles de fenêtre glissante par la manière dont elles génèrent et traitent ces régions avant la phase d'extraction de caractéristiques et de classification au sein d'un réseau neuronal convolutif (CNN). R-CNN [109], l'une des méthodes pionnières de cette catégorie de réseau de détection, remplace la fenêtre glissante par un algorithme externe (Selective Search) précédant le réseau convolution afin de réduire le nombre de régions d'intérêt à 2000 en se basant sur les éléments visuels saillants. Cependant, une limitation majeure de R-CNN réside dans son besoin d'extraire et de traiter les caractéristiques de chaque région individuellement, ce qui induit une complexité computationnelle considérable. Une version évoluée de R-CNN nommée Fast R-CNN [110] modifie la structure d'extraction de caractéristiques sur l'ensemble de l'image puis une projection des ROIs prédites sur cette carte à partir d'une couche de

regroupement pour normaliser les dimensions de chaque ROI, indépendamment de leurs dimensions initiales, pour produire une représentation de dimensions fixes. Cependant, un goulot d'étranglement ou *bottleneck* entre la prédiction des ROI et l'extraction des caractéristiques subistes. Faster R-CNN [111] est l'une des approches par région proposée les plus couramment utilisée pour la détection d'objet au sein d'une image 2D. Cette méthode corrige les défauts des anciennes méthodes pour permettre un fonctionnement temps-réel requis par de nombreuses applications modernes tel que les méthodes d'assistance pour personnes malvoyantes. L'extraction des ROI par un algorithme externe est remplacée par un module dédié (RPN - Region Proposal Network) qui élimine ainsi la nécessité de recourir à un algorithme externe pour la proposition des ROI. Cette intégration transforme le modèle en un système end-to-end, optimisant ainsi le processus de détection.

La proposition de régions suit plusieurs étapes pour générer une région pertinente et associer une classe à cette région extraite bien que Faster R-CNN accélèrent ces processus. Des méthodes plus récentes condensent ces multiples étapes de génération de ROI jusqu'à la classification en une seule, utilisant un unique réseau CNN basé sur des techniques de régression comme les architectures SSD (Single Shot Detector) et YOLO (You Only Look Once) qui représente les réseaux de détection les plus répandus dans les applications temps réel. SSD est le premier réseau de régression à atteindre les performances des méthodes à deux étages en reposant sur un backbone dérivé de VGG-16 avec des convolutions additionnelles ajoutées en fin du réseau. Ces couches supplémentaires voient leurs dimensions décroître graduellement, formant une structure pyramidale afin de détecter des objets de différentes tailles. Les objets de petite taille sont identifiés dans les premières couches où les détails de l'image sont préservés, tandis que les couches plus profondes détectent les objets de plus grande taille. YOLO est une famille de réseau de détection partageant des propriétés communes comme une unique traversée de l'information visuelle pour réaliser ses prédictions, d'où son nom "You Only Look Once" à l'opposé des méthodes à région proposée. Cette architecture vise à atteindre des performances élevées sans compromettre le temps d'exécution. En effet, la première version du modèle permet d'atteindre un processus en temps réel avec cependant des faiblesses pour distinguer les objets de petite taille [112]. Au fil du temps, diverses évolutions et adaptations de l'architecture originale ont été introduites afin d'améliorer la précision et les performances du modèle. Cependant, les versions postérieures à YOLOv4 divergent en termes d'architecture par rapport à la conception initiale, mais conservant l'essence fondamentale de l'exécution temps réel des versions antérieures.

Les performances des méthodes de détection d'objets par boîtes englobantes sont évaluées à partir de métriques spécifiques pour l'indice de classification et les

coordonnées de la boîte sur l'image 2D. L'exactitude de la classification suit les mêmes métriques que pour les méthodes de classification évoquées précédemment, tandis que les coordonnées sont comparées par une métrique nommée intersection sur l'union (IoU ∼ annexe 7.3) qui recherche le taux de recouvrement entre les valeurs exactes et celles prédites. L'évaluation de l'IoU permet de s'assurer que le modèle détecte non seulement l'objet correctement, mais aussi qu'il le localise avec précision.

2.3.2.3 Segmentation sémantique

Les méthodes de segmentation sémantique basées sur des CNN s'inscrivent dans la continuité des approches de classification. Toutefois, au lieu d'attribuer une étiquette unique à l'ensemble de l'image, elles assignent une étiquette spécifique à chaque pixel, cette tâche est communément appelée prédiction dense. La figure 2.23 représente la segmentation d'une image en six classes (arbre, route, véhicule, personne, ciel, poteau). Cette granularité accrue permet d'obtenir des informations précises sur la nature sémantique de chaque pixel en tenant compte de son contexte spatial. Une fois regroupés, les pixels contigus partageant une étiquette similaires délimitent la forme et l'emplacement d'un objet ou d'une entité spécifique au sein de l'image. Les méthodes de segmentation sémantique par CNN s'organisent avec une architecture similaire aux méthodes de classification par deep-learning évoquée précédemment avec un réseau d'extraction et d'encodage des informations caractéristiques. Cependant, au lieu de définir une seule classe pour l'ensemble des informations en réduisant le nombre de caractéristiques, les dimensions des cartes de caractéristiques sont ensuite augmentées afin de correspondre aux dimensions de l'image d'origine. Une architecture populaire dans les modèles de segmentation d'images est basée sur une structure d'encodeur suivi d'un décodeur.

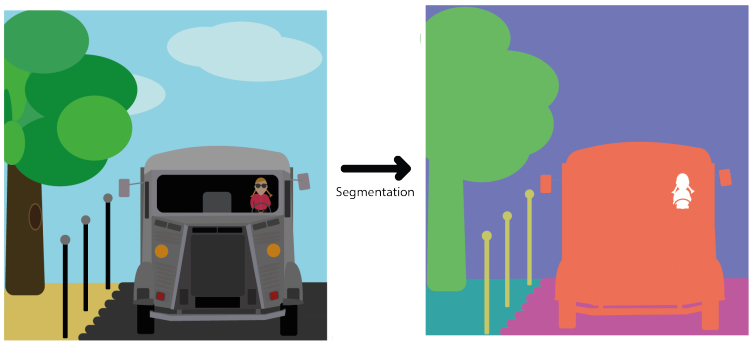


FIGURE 2.23 – Segmentation sémantique d'une image.

Dans la phase d'encodeur, l'image est progressivement sous-échantillonnée pour

obtenir une résolution spatiale réduite. Cette réduction permet de créer des cartes d'entités à basse résolution, mais très discriminantes pour distinguer les différentes classes d'objets présentes dans l'image. Le décodeur réalise la tâche inverse du décodeur avec un suréchantillonnage des cartes d'entités précédemment obtenues afin de générer une carte de segmentation de haute résolution. Il existe plusieurs approches pour augmenter la résolution d'une carte de caractéristiques telles que les opérations de dégroupement ou des convolutions transposées. Les opérations de dégroupement (*unpooling*) réalisent l'opération inverse aux méthodes de regroupement où la résolution spatiale est augmentée en distribuant une seule valeur vers une résolution plus élevée. Les convolutions transposées, largement répandues, sont des méthodes de suréchantillonnage avec paramètres où une valeur d'une matrice de basse résolution et multipliés par l'ensemble des poids du noyau puis projetée dans une matrice de résolution supérieure. U-Net [113] est l'un des modèles les plus utilisés dans des taches de segmentation où des connexions lient directement les couches de l'encodeur et celle du décodeur de même résolution spatiale. Ces connexions ont pour effet de transférer des informations spatiales détaillées de l'encodeur au décodeur, facilitant ainsi la reconstruction d'objets avec des formes précises lors du processus de suréchantillonnage. En termes d'architecture, l'U-Net est structuré comme un "U" avec un chemin de contraction (l'encodeur) qui capte le contexte global de l'image, suivi d'un chemin d'expansion symétrique (le décodeur) pour une localisation fine des objets. D'autres variantes de U-Net remplacent des couches de convolutions par des couches denses ou résiduelles, tandis que d'autres architectures ont été proposées pour accroire les performances.

2.3.3 Application aux méthodes assistances à la navigation pour les personnes malvoyantes

Les systèmes d'assistance à la navigation destinés aux personnes malvoyantes exploitent les techniques de vision par ordinateur mentionnées précédemment afin d'augmenter les perceptions naturelles avec des informations additionnelles du champ visuo-spatial non perceptible. La figure 2.24 illustre le fonctionnement d'un système d'assistance à la navigation combinant les deux aspects essentiels pour une navigation sûre dans un espace non familier que sont la détection des obstacles (ETA) et la direction à suivre (EOA). Des informations annexes à celles extraites dans la scène visuelle, tels que des données de localisation (GPS, GSM, Balise bluetooth) [114] peuvent être associées en fonction des contraintes du milieu de navigation lorsque l'information visuelle n'est pas

suffisante pour définir la trajectoire.

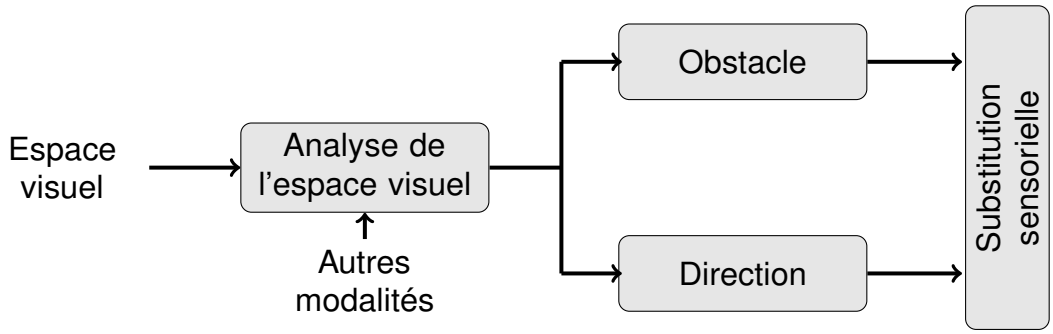


FIGURE 2.24 – Méthode de navigation d'assistance pour personne déficiente visuelle.

La reconnaissance d'éléments essentiels, tels que les obstacles, est fondamentale pour toutes les méthodes ETA afin d'assurer la sécurité des personnes déficientes visuelles. Ces méthodes d'assistances essaient de simuler le comportement naturel d'une personne voyante lors de ces déplacements pour définir la présence d'obstacles à partir de critères tels que leur nature, leur localisation dans l'espace ou leur vitesse de déplacement. L'identification des éléments d'une scène visuelle se fait au moyen de techniques de classification par ML ou apprentissage profond en basant leurs phases d'apprentissage sur des objets typiques de l'environnement de déplacement, par exemple une porte dans un couloir ou un arbre dans un parc. Third-eyes [115] cherche à identifier des aliments dans un magasin à l'aide de deux caméras : l'une fixée sur des lunettes identifiant l'aspect général de l'objet comme la présence d'une boite de céréales et une deuxième sur des gants, plus proches des aliments, permettant de distinguer deux packagings différents. La méthode d'identification combine l'extraction des descripteurs via SURF avec une classification SVM. Des techniques comparables emploient les descripteurs HOG ou SIFT, associés à des algorithmes de classification comme K-Means ou des méthodes d'apprentissage profond. Dans le cadre du système Cross-Safe [116] qui assiste les malvoyants à traverser les routes en classifiant les feux piétons aux intersections grâce à un CNN compact de 8 couches (4 couches de convolutions, 2 couches de regroupement, 2 couches denses). Des réseaux sont utilisés pour accéder à des objets de natures multiples, par exemple la présence de 11 éléments prédéfinis symbolisant les objets les plus répandus dans un environnement extérieur tel que les voitures ou les piétons à partir d'une structure VGG19 [117] ou pour identifier les objets communs dans une structure hospitalière [118]. D'autres architectures modernes et légères existent, dédiées pour des systèmes embarqués avec capacités de calcul restreintes comme Resnet [119], Mobilenetv2 [120] ou EfficientNet [121].

La connaissance de la présence d'un élément dans un espace peut s'avérer être limitée sans la connaissance de sa position. La localisation des éléments pertinents est réalisée par des approches de subdivision de la scène visuelle en grille simple [116] ou à l'aide de réseaux de convolutions dédiés à la localisation comme les réseaux de détection d'objet conçus pour des applications temps réel afin d'émettre rapidement une information sensorielle d'alerte tel que les architectures SSD [122] ou YOLO [100], [123]. L'intégration d'informations de profondeur enrichit la représentation 2D en y ajoutant une dimension supplémentaire, rendue possible grâce à l'utilisation de caméras de profondeur [124] ou de LIDAR [125]. Par exemple, certains systèmes [126], [127] exploitent un seuillage sur une image de profondeur pour alerter l'utilisateur de la présence d'un obstacle proche. La détection d'éléments mobiles dans l'environnement, en particulier ceux se déplaçant rapidement en direction d'une personne, représentent un danger important. La simple prise en compte de la proximité ne suffit pas toujours pour anticiper ces éléments et les distinguer des objets immobiles. Leur identification est réalisée sur un ensemble d'images successives où les éléments identifiés sur une image. Le suivi de ces dernières est réalisé sur les images subséquentes en établissant des correspondances entre les informations d'une image à l'autre à partir de méthodes d'appariement de descripteurs les plus similaires. [128] propose une technique d'appariement basée sur les boîtes englobantes en utilisant l'algorithme hongrois, associé à une méthode de prédiction comme le filtre de Kalman, appelée SORT, pour déduire le mouvement des objets.

En complément des informations d'obstacles proposées par les systèmes ETA, les systèmes d'orientation et d'assistance (EOA) aident à diriger une personne malvoyante à travers son environnement pour atteindre une destination spécifique. En milieu urbain ou à l'intérieur d'un bâtiment, les trajectoires piétonnes sont souvent dictées par la configuration topologique des lieux, comme la disposition des murs ou un chemin tracé dans un parc [129]. Des méthodes de segmentation de l'image sont employées soit à partir d'information caractéristique comme la couleur saillante d'un chemin podotactile ou pour discriminer un espace libre sur le sol à partir d'une méthode de segmentation planaire basée sur des nuages de points [130], [131] ou par segmentation sémantique [129] d'une image l'aide de réseau de neurone. Ainsi, un système d'assistance [132] utilise un CNN pour réaliser une segmentation sémantique à partir d'images RGB-D. Une fois l'image segmentée, elle est ensuite soumise à un réseau de neurones de classification, également basé sur un CNN, qui détermine l'itinéraire à emprunter pour contourner un obstacle ou pour suivre le long d'un mur.

2.3.3.1 Jeux de données pour la navigation

Une généralisation précise d'un phénomène ou de la détection d'un élément pertinente comme un obstacle dans une scène visuelle est fortement corrélée à la quantité et à la qualité des données au cœur des méthodes de ML et des réseaux d'apprentissage profonds. Un ensemble de données inadéquat ou incomplet pour modéliser une tâche spécifique peut mener à des résultats incorrects. Les systèmes d'assistance aux personnes malvoyantes, visant à renforcer la perception et la compréhension de l'environnement, qu'ils agissent d'un espace intérieur ou extérieur de navigation, s'appuient sur des données spécifiquement conçues pour décrire ces environnements, comme illustré dans le tableau 2.2. Chaque jeu de données à ses propres caractéristiques, que ce soit en termes d'annotations, de variété d'images ou des angles de vue proposés. Ainsi, les bases d'images des ensembles de 1 à 6 sont orientées vers la détection d'obstacles. Parmi ces 6 jeux de données, MS COCO et Pascal, sont des bases globales servent de référence pour l'évaluation, en raison de la diversité des lieux d'acquisition, aussi bien intérieur, d'extérieur, et des objets présents. Les jeux de données suivants sont réservés à des applications de segmentation sémantique. Un jeu de données nommé SideGuide [133] (Détection d'objet : 6 et segmentation sémantique : 16) a été élaboré dans le but précis de soutenir le développement de systèmes d'assistance dédiés aux personnes aveugles par l'apport d'informations sémantiques sur l'espace de navigation. Cette base de données comporte des annotations relatives au revêtement du sol, en précisant sa nature. De plus, elles incluent des annotations concernant 24 objets couramment retrouvés dans les zones piétonnes en Corée du Sud. D'autres bases, initialement conçues pour les véhicules autonomes (11-14 et partiellement 15), présentent néanmoins des caractéristiques pertinentes pour les zones piétonnières et routières. Bien que SideGuide représente un atout majeur dans l'élaboration des méthodes d'assistances visuelles, ce jeu de donnée intègre un nombre limité de scènes capturées et d'objet caractéristiques constituant un espace de navigation extérieur. Cette limitation est accentuée dans le cadre d'une navigation dans un environnement urbain où la forme du mobilier ainsi que les éléments structurant le plus ces espaces diffèrent. De plus, l'absence de séquences d'information contraint l'analyse de l'image sans tenir compte de l'information temporel pertinente lors d'un déplacement. Un jeu de données plus ouvert à des espaces et des informations complémentaires, voire constitué de séquences vidéo, serait utile pour améliorer le développement de méthodes mieux adaptées à des environnements variés.

	Nom	Images	Type	Nom. de classes	Annotations	Environnement	Ref.
1	MS COCO	328000	RGB	80	Détection	Intérieur Extérieur	[134]
2	PASCAL	11540	RGB	20	Détection	Intérieur Extérieur	[109]
3	BDD100K	100000 vidéos	RGB	10	Détection	Extérieur	[135]
4	A2D2	12497	RGB	10	Détection 3D	Extérieur	[136]
5	SUN	11540	RGB	20	Détection	Extérieur	[137]
6	SideGuide	25000	RGB	66	Détection	Extérieur	[133]
7	ADE20K	25000	RGB	150	Segmentation	Intérieur Extérieur	[138]
8	NYUv2	1449	RGB-D	30	Segmentation	Intérieur	[49]
9	Scannet	5000	RGB-D	30	Segmentation	Intérieur	[139]
10	MS COCO Stuff	164000	RGB	172	Segmentation	Intérieur Extérieur	[140]
11	A2D2	12497	RGB	38	Segmentation	Extérieur	[136]
12	ApolloScape	140000	RGB	28	Segmentation	Extérieur	[141]
13	BDD100K	10000 vidéos	RGB	40	Segmentation	Extérieur	[135]
14	Cityscape	5000	RGB-D	30	Segmentation	Extérieur	[142]
15	Mapillary Vistas	25000	RGB	66	Segmentation	Extérieur	[143]
16	SideGuide	5000	RGB	66	Segmentation	Extérieur	[133]

TABLE 2.2 – Liste des jeux de données de segmentation et de détection d'environnement intérieur et extérieur.

2.4 Limitations des systèmes actuels et objectif de la thèse

Les personnes malvoyantes font face à de multiples problèmes lors de leurs déplacements comme traverser une route, atteindre un lieu sans connaissance mentale préalable. Des outils technologiques ont été développés pour les accompagner dans leurs quotidiens et pallier la déficiente d'information visuelle à l'aide de stimulus sensoriels complémentaires les renseignant sur leur environnement. Néanmoins, ces outils et méthodes d'assistance doivent répondre à des enjeux spécifiques, tant liés aux besoins intrinsèques des utilisateurs qu'aux contraintes opérationnelles de ces dispositifs. Une analyse approfondie des solutions d'assistance visuelle établit des critères de performance, catégorisés en neuf domaines [144] :

- Type et précision des capteurs
- Période de fonctionnement
- Environnement de fonctionnement
- Temps de réponse
- Qualité des retours sensoriels
- Portée de détection
- Ergonomie
- Robustesse des traitements
- Coût du système

Sur la base de ces critères, des évaluations subjectives, pondérées selon leur pertinence, sont proposées pour définir la performance globale des systèmes visant à pallier l'absence de métriques objectives directement quantifiables [145]. Une solution idéale serait capable de fonctionner efficacement en journée comme en nuit, offrant des informations sensorielles pertinentes et adaptées à chaque situation, tout en étant robuste et ergonomique. Cependant, l'ensemble de ces facteurs sont parfois antinomiques comme le besoin d'une analyse visuelle détaillée et rapide tout en maintenant le dispositif peu volumineux et économe en énergie. Pour répondre à ces contraintes, des méthodes [146]-[148] font appel à des technologies de cloud computing pour le traitement des informations

coûteuses en ressources matérielles, mais en s'exposant à une rupture du signal de communication, qu'il soit par téléphonie mobile (GSM, GPRS, LTE, 5G) ou par d'autres réseaux de communication (LoRa, SigFox, etc). De plus, les systèmes actuels d'assistance à la navigation tendent à se focaliser soit sur la détection des obstacles, soit sur la détermination du chemin à suivre, sans nécessairement intégrer de manière simultanée ces deux dimensions pourtant cruciales à la sécurité de la navigation. L'agrégation harmonieuse de ces attributs est indispensable pour assurer une mobilité sûre dans un espace non familier par une personne aveugle. De surcroît, la prise en compte des informations spatiales pertinentes pour une navigation sécurisée demeure imparfaite. Par exemple, des critères tels que la sélection d'itinéraires minimisant les risques d'accidents ou l'identification d'éléments distinctifs d'un environnement urbain, comme la démarcation entre zones piétonnes et voies de circulation motorisée, ne sont pas systématiquement intégrés dans le processus de détermination de la trajectoire.

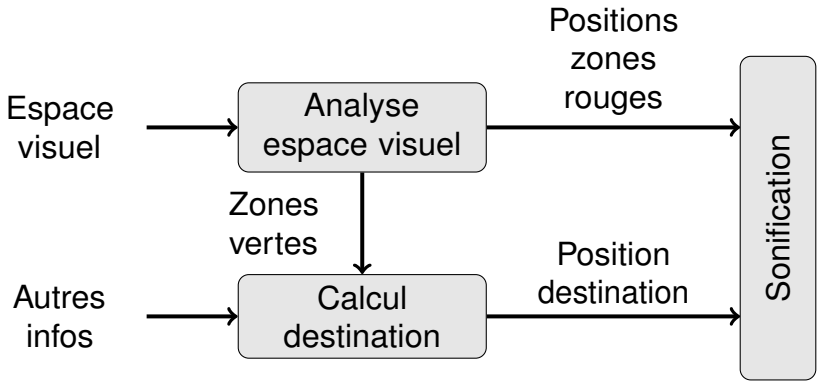


FIGURE 2.25 – Processus de détermination des zones de navigation accessible ou dangereuse.

Les travaux présentés dans ce manuscrit, motivés par les capacités d'analyse d'information visuelle par ordinateur accrues et par les hautes facultés de localisation de signaux sonores, visent au développement d'un dispositif d'assistance à la mobilité pour les personnes malvoyantes. En effet, nous proposons d'enrichir les informations perçues par une personne malvoyante à l'aide d'un système temps-réel de puissance modérée basé sur une substitution sensorielle visuelle vers auditive. En effet, la modalité auditive permet une grande variété de personnalisation de signaux auditifs compréhensible par le cerveau, enveloppant l'émission d'informations verbales et des sons spatialisés stéréophoniques. Leur nature contraire, aussi bien dans leur durée d'émission que par leur propriété informative, offre une vaste palette de signaux sonores œuvrant de concert pour s'adapter au contexte d'une situation rencontrée. Une information verbale a un caractère plus sémantique tandis qu'un son spatialisé stéréophonique indique avec rapidité et précision le positionnement d'un élément marquant et crucial (section 2.2). Ainsi, au sein

de notre méthode, nous proposons de distinguer par de brefs sons spatialisés différents les informations relatives au chemin à suivre et aux obstacles environnants. D'autre part, les informations visuelles acquises par un système d'acquisition et analysées par des traitements d'images modernes par réseaux de neurones convolutifs permettent d'accéder à une connaissance enrichie de l'espace de navigation tout en limitant la charge de calcul par des voies d'optimisations matérielles et logicielles. L'association d'une part d'un réseau de neurone dédié à une application de segmentation d'image et d'autre part d'un autre réseau de détection d'obstacle permet de définir des zones sur lesquelles la circulation est libre de tout obstacle, dénommées "zones vertes" et d'autres parts les espaces présentant des dangers, appelés "zones rouges" (figure 4.1). La figure 2.25 illustre cette segmentation, mettant en évidence le processus de distinction de ces zones et leurs implications dans le signal sonore. Dans un objectif de proposer un système réaliste pour une utilisation du système, nous proposons une méthode autonome ne nécessitant aucune ressource de calcul ou de communication externe au système autonome proposé afin d'éviter les problèmes associés. Cette approche est guidée par l'intégration d'une méthode d'encodage de l'information visuelle vers sonore avec une faible latence et des techniques de vision artificielle temps réel pour garantir une synchronisation parfaite entre les deux représentations spatiales. De plus, l'approche proposée conserve un fonctionnement sur une ressource matérielle de consommation énergétique modérée pour permettre l'autonomie du système. En effet, des techniques de vision artificielle permettent une analyse à la fois efficace et robuste, tout en conférant des facultés d'interprétation de données traditionnellement négligées par les approches standards. Elles se distinguent notamment par leur aptitude à identifier et à différencier des zones distinctes, telles que la séparation entre espaces piétonniers et voies routières. Au cours des prochains chapitres, il sera abordé les différents points cruciaux relatifs à ce système d'assistance à la navigation tel que le processus d'encodage de l'information visuelle en signaux sonore, l'analyse de l'espace visuel et la méthode de navigation. Cette dernière permettant à une personne aveugle de se mouvoir dans un espace urbain en toute sécurité sans heurter un obstacle jusqu'à atteindre une destination.

3

Interprétation d'une scène visuelle dans un espace sonore

Sommaire

L'encodage de l'information visuelle en signaux sonores est la pierre angulaire de tous les systèmes de substitution de la vision par l'audition (SVA) destinés aux personnes malvoyantes. Le principe est de transformer l'information visuo-spatiale en information auditive. La transcription d'information visuelle vers une information auditives s'effectue principalement au moyen d'expressions sémantiques orales qui décrivent la situation comme le ferait un être humain voyant, mais également par l'émission de sons spatialisés. Bien que ces sons puissent sembler dénués de sens au début de l'utilisation de tels systèmes, ils sont le moyen de véhiculer un volume important d'informations spatiales permettant la réalisation de tâches de navigation. Ces informations nécessitent pourtant une phase d'apprentissage et de familiarisation pour être correctement interprétées. Cependant, les capacités réceptives auditives ne sont pas aussi vastes que celles de la modalité visuelle. Le tableau 2.1, expliqué précédemment, illustre les faiblesse des modalités sensorielles, à l'exception de la vision, pour capter une information spatiale. La modalité auditive, en particulier, peut être facilement perturbée lorsque plusieurs sources sonores se manifestent simultanément, entraînant des interférences entre les différents signaux reçus. Ces interférences peuvent compromettre la précision de la perception auditive. Sur le plan neurologique, l'intégration excessive d'informations sonores simultanées peut conduire à une saturation des circuits cérébraux impliqués dans le traitement auditif. Cette surcharge cognitive engendre une diminution de la capacité à isoler et à traiter de manière distincte et efficace chaque source d'information sonore.

Les informations sonores transmissent pour assister une personne malvoyante, avec une méthode de SVA, doivent ainsi être particulièrement claires, précises, adaptées à leurs besoins et se restreindre aux éléments les plus critiques pour l'accomplissement d'une tâche demandée sans gêner leurs perceptions naturelles. Les informations à transmettre varient en fonction de la tâche à accomplir. Dans le contexte de l'assistance à la navigation, deux informations sont fondamentales pour parvenir à une destination souhaitée. La première concerne la direction à suivre pour atteindre un lieu souhaité, tandis que la seconde intègre la présence des obstacles environnants entravant la progression de la personne et présentant un danger important s'ils ne sont pas perçus avec précision et rapidité par l'utilisateur. L'exploitation des techniques de spatialisation sonore stéréophonique permet une localisation précise et rapide d'informations tridimensionnelles. En cas de génération de stimulus sonores longs, un déphasage peut se produire entre la localisation véritable de l'élément et celle ayant servi lors de la synthèse sonore, notamment lors de mouvements rapides de l'utilisateur ou des éléments au sein de la scène visuelle. Bien que la détection de tous les obstacles soit utile au déplacement d'une personne aveugle, leurs dangerosités ne sont pas équivalentes. Les objets proches présentent un danger bien

plus immédiat, et leur emplacement doit être connu en priorité par l'utilisateur. À l'opposé, les objets plus lointains pourront devenir potentiellement dangereux lorsque l'utilisateur s'en approchera. Par conséquent, une hiérarchisation des informations visant à réduire le nombre d'informations transmises peut se révéler nécessaire.

Ce chapitre se concentrera sur l'interaction entre une personne aveugle et un système de substitution sensorielle d'assistance à la navigation, notamment sur la capacité de la personne à interpréter des stimulus auditifs représentant une information visuelle. Tout d'abord, le processus d'encodage d'une information visuelle en information sonore sera abordé. Nous présenterons ensuite l'extension de cette méthode à un ensemble plus vaste d'information de localisation tel que celui utilisé pour le signalement de multiples obstacles proches. Cette partie sera accompagnée par une étude des capacités humaines à localiser précisément un stimulus, qu'il soit visuel ou auditif. Enfin, une description de la méthodologie de fusion des deux informations fondamentales à la réussite d'une tache de navigation sera traitée puis évaluée à partir d'une preuve de concept d'assistance à la mobilité dans un bâtiment.

3.1 Représentation d'une information visuelle dans un espace auditif

L'encodage d'une information visuelle ponctuelle en un signal acoustique adapté à une tâche spécifique est une problématique centrale. Une faible précision de localisation sonore peut réduire grandement l'intérêt des informations encodées et par conséquent l'intérêt d'un système d'assistance de substitution vision-audition. De plus, les signaux émis peuvent être préjudiciables pour un utilisateur non-voyant si ceux-ci le guident vers une zone dangereuse ou contredisent les informations à sa disposition. Un protocole d'encodage visuo-sonore doit donc être suffisamment précis et facilement intelligible par l'utilisateur. De surcroît, dans le cadre d'un système d'assistance à la navigation , et comme évoqué précédemment en ouverture de ce chapitre, la localisation d'un élément doit pouvoir être réalisé dans un temps réduit à travers une émission sonore courte.

Les méthodes d'encodage visuo-auditif présentées dans la section 2.2.2 introduisaient deux familles d'encodages : par balayage temporel progressif ou par transmission globale. La première catégorie repose sur une transmission séquentielle d'une information élément par élément tandis que la seconde, en revanche, vise à transmettre l'ensemble des informations de manière simultanée. Bien que les deux méthodes aient chacune leurs mérites et permettent de détecter et de localiser avec précision un objet dans l'espace,

la méthode de balayage temporel présente un inconvénient majeur : son processus séquentiel limite son efficacité dans des situations d'urgence où une réaction immédiate est nécessaire pour éviter un danger. Par conséquent, la méthode d'encodage des éléments proches utilisée dans nos travaux s'est orientée vers une méthode globale. Cependant, ces méthodes globales faisant transiter en simultané une multitude d'informations spatiales, une combinaison méticuleuse des indices acoustiques est nécessaire pour rendre l'interprétation du signal sonore plus accessible.

3.1.1 Encodage d'une coordonnée azimutale

Les capacités auditives pour localiser l'origine d'un bruit offrent des fondations à un système d'encodage des informations visuo-spatiales dans un espace sonore. Cette capacité repose principalement sur l'analyse par le système cognitif des différences spectrales et temporelles entre les signaux perçus par l'oreille gauche et l'oreille droite, différences principalement constituées par ce que l'on appelle la différence d'intensité (ILD inter-aural level difference) et la différence temporelle (ITD inter-aural time difference). La transformation d'un signal sonore monophonique en une paire de signaux stéréophoniques spatialisées consiste principalement à reproduire ces indices acoustiques. Des fonctions de transferts relatives à la tête, nommées HRTF, visent à simuler les différences de temps et d'intensité que l'oreille humaine exploite naturellement pour localiser un son. Ce mécanisme, illustré par les équations 3.1 et 3.2, transforme un signal monophonique classique, nommé *m(t)*, en signaux stéréophoniques à l'aide d'une convolution avec une paire de HRTF reproduisant un signal reçu par le conduit auditif à un angle azimutal φ d'une personne. L'usage de fonctions HRTF bidimensionnels intégrant l'angle d'élévation améliore la spatialisation de l'information sonore .

$$s_g(\varphi, t) = HRTF_g(\varphi) * m(t) \tag{3.1}$$

$$s_d(\varphi, t) = HRTF_d(\varphi) * m(t) \tag{3.2}$$

L'encodage sonore pour la reproduction de l'angle azimutal est crucial dans la conception de systèmes de substitution sensorielle vision-audition. Chaque signal monophonique est convolué par des HRTF issues d'enregistrements réalisés sur un mannequin. Pour ce faire, nous avons utilisé la base de données CIPIC [149] qui contient les HRTF d'un mannequin de type KEMAR avec une source sonore placée sous différents angles d'azimut et d'élévation, avec des intervalles respectifs de 5° et 5,625°. Cependant, la

quantité restreinte d'HRTF disponibles dans la base de données ne permet pas de couvrir l'ensemble des positions angulaires possibles. Une interpolation spatiale à quatre points est mise en œuvre pour calculer les valeurs intermédiaires, palliant ainsi les lacunes de la base de données en termes de diversité angulaire. Durant ce processus, la Différence de Niveau Interaurale (ILD) et les signaux issus de la convolution sont interpolés séparément, avant d'être réassemblés [150]. La variation d'amplitude entre les signaux sonores stéréophoniques est illustrée dans les figures 3.1 et 3.2 pour encoder une information située à gauche et à droite d'une personne avec un angle de 40°.

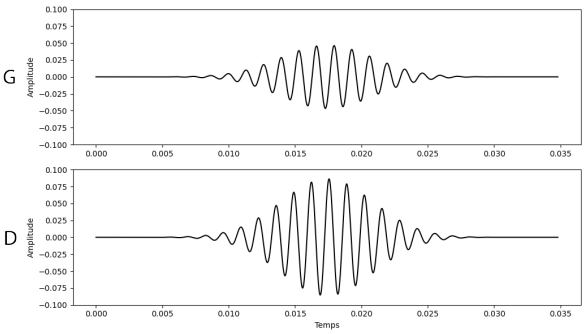


FIGURE 3.1 – Signal sonore d'une information spatiale avec un angle azimutal de 40° (droite) perçu par l'oreille gauche (haut) et droite (bas).

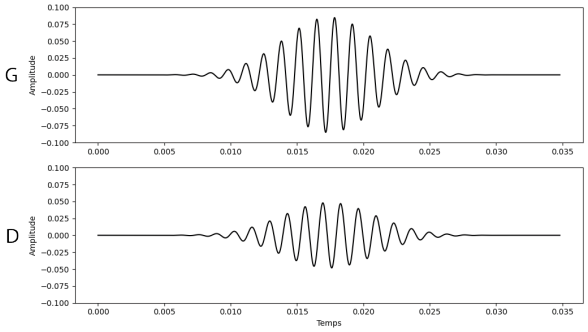


FIGURE 3.2 – Signal sonore d'une information spatiale avec un angle azimutal de -40° (gauche) perçu par l'oreille gauche (haut) et droite (bas).

3.1.2 Encodage de la composante verticale

La capacité humaine à déterminer la position d'origine d'un son varie selon qu'il s'agit de localiser sa position verticale ou horizontale. Alors que l'orientation azimutale (plan longitudinal) est généralement distinguée sans difficulté majeure, la détection de l'élévation (plan médian) d'une source sonore est sujette à plus de confusions. Les indices acoustiques nécessaires pour discerner l'élévation sont principalement issus de masquages fréquentiels réalisés par la réverbération de l'onde sonore sur le pavillon de l'oreille. Ces indices sont donc plus difficiles à reproduire puisqu'ils dépendent de la morphologie des oreilles se chacun. Un deuxième indice acoustique en complément de la spatialisation par HRTF peut alors être employé en vue d'améliorer la précision de la localisation. L'intégration de cet indice acoustique complémentaire peut s'appuyer sur les composantes spectrales des sons générés comme la hauteur.

$$tone(\theta, t) = \sin(2\pi \cdot freq(\theta) \cdot t + \psi) \tag{3.3}$$

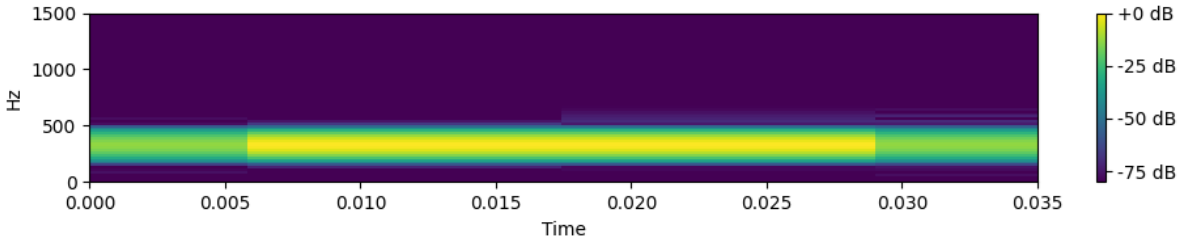


FIGURE 3.3 – Encodage sonore d'une d'information spatiale avec un angle en élévation de -20°.

Dans cette optique, l'information sonore spatialisée à l'aide d'une HRTF est enrichie par l'intégration d'un indice fréquentiel dans le signal sonore. Ainsi, le son monophonique original, *m(t)* est modifié pour que sa hauteur tonale, corresponde à l'élévation de la source sonore. En d'autres termes, cette hauteur est modifiée pour refléter l'élévation dans le champ de vision de la caméra du détail graphique que l'on cherche à encoder. Ce signal est défini dans 3.3 avec une phase variable ψ . L'amplitude de l'élévation est représentée par une échelle linéaire de mels, allant de 344 mels pour les positions basses à 1286 mels pour les hautes. Cette gamme couvre approximativement des fréquences de 250 Hz à 1492 Hz, avec 120 niveaux intermédiaires. Les spectrogrammes des signaux sonores pour deux informations visuelles situées à des positions relatives de 20° et -20° sont illustrés dans des figures 3.3 et 3.4. D'autres méthodes utilisent d'autres échelles fréquentielles, mais nous avons préféré rester sur celle-ci puisqu'une gamme de fréquence plus large peut devenir plus désagréable à l'écoute.

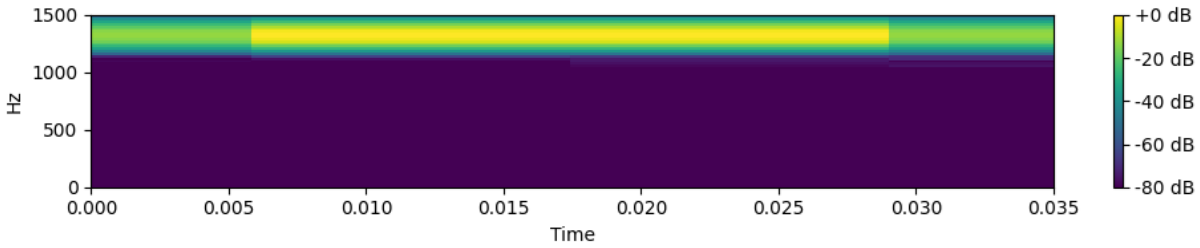


FIGURE 3.4 – Encodage sonore d'une d'information spatiale avec un angle en élévation de 20°.

3.1.3 Encodage de la distance

Naturellement, la distance qui sépare un auditeur d'un bruit est estimée à partir de l'atténuation de son intensité en fonction de la distance parcourue. Ce phénomène s'explique physiquement par l'augmentation de la surface à travers laquelle l'onde sonore se déplace. Une transformation de l'énergie sonore en énergie thermique se produit

également. Cette atténuation acoustique est souvent modélisée par une réduction de l'intensité sonore proportionnelle au carré de la distance parcourue. Un indice acoustique associé à l'intensité du signal permet de simuler la propagation d'une onde sonore éloignée.

$$s_g(\varphi, \theta, z, t) = A(z) \cdot HRTF_g(\varphi, \theta) * tone(\theta, t) \tag{3.4}$$

$$s_d(\varphi, \theta, z, t) = A(z) \cdot HRTF_d(\varphi, \theta) * tone(\theta, t) \tag{3.5}$$

Le système d'encodage 2D, établi précédemment pour traiter les informations verticales et horizontales, est alors enrichi avec un facteur encodant la profondeur par l'intensité sonore. Les équations 3.4 et 3.5 illustrent l'encodage d'une position tridimensionnelle dans l'espace visuo-spatial codée dans un espace sonore relatif. Le processus intègre une fonction de modulation de l'intensité sonore par rapport à une distance, noté A. Ce mécanisme produit une série de signaux sonores associés à différentes positions d'une durée totale de 33 ms. La durée est définie par la fréquence d'acquisition de la caméra fixée à 30 images par seconde. Cette série inclut une augmentation progressive de l'intensité suivant une courbe en cosinus pendant les cinq premières millisecondes, et une diminution identique sur les cinq dernières millisecondes. Cette atténuation du signal en entrée et en sortie favorise une transition sonore douce et minimise les artefacts potentiels entre différentes images encodées.

3.2 Étude des capacités humaines de localisation d'un élément dans l'espace auditif

Les capacités humaines de localisation sonore sont limitées tant sur le plan perceptif que cognitif lorsqu'il s'agit d'interpréter avec précision la position de multiples sons dans un environnement spatial [151]. Les inexactitudes dans ce processus d'interprétation peuvent réduire significativement l'utilité des informations transmises. En effet, un décalage angulaire important entre la position réelle et celle interprétée peut conduire à la réalisation d'une action inadéquate. Les limites de performances de localisations d'objets obtenues par un encodage audio-spatial doivent donc être mesurées, tant sur le plan horizontal (azimutal) que vertical (élévation).

3.2.1 Performance en localisation azimutale

Nous avons réalisé une étude sur l'évaluation des capacités humaines en termes de perception et de localisation azimutale, en s'appuyant sur le processus d'encodage décrit précédemment. L'objectif principal est de quantifier la précision avec laquelle des individus peuvent déterminer la position angulaire azimutale d'un élément visuel dans un environnement auditif. À cette fin, un protocole expérimental a été monté dans lequel un participant devait estimer la position azimutale d'une personne debout dans son environnement proche. Le participant assis sur une chaise pivotante était positionnée au centre d'un cercle de 4 mètres de diamètre tandis que la personne cible se plaçait de manière aléatoire sur l'une des huit positions prédéfinies. Ces positions étaient équitablement espacées avec un intervalle angulaire fixe de 45°. Le participant devait pivoter sur sa chaise afin d'orienter sa tête le plus précisément possible vers la personne cible, puis valider son estimation de la position en cliquant sur un bouton. Cette estimation était par la suite comparée à la position réelle de la personne en vue de définir l'erreur angulaire. Ces configurations sont illustrées et décrites plus précisément dans la figure 3.5. L'évaluation fut réalisée sous deux conditions expérimentales. La première avec une assistance auditive et les yeux des participants bandés. Puis sous une seconde condition de contrôle reflétant les capacités visuelles naturelles d'une personne non entravée par un dispositif ou un bandeau au niveau des yeux.

Le dispositif expérimental utilisé était composé de deux traqueurs de positions HTC Vive, dont l'un sur la tête de la personne aveugle et le second sur le sternum de la personne cible, et d'un système de substitution visuo-auditif. La position de la personne cible servant de base au signal sonore généré par le système était estimée par un algorithme produit sur la base d'une méthode de détection d'objet par apprentissage profond détaillée dans le chapitre 4. La position spatiale encodée en un signal sonore est définie à partir de la coordonnée horizontale du centroïde de la boîte englobant la personne cible.

Dans le cadre de cette expérience, nous avons sélectionné 10 participants sans imposer de contraintes particulières quant à leur âge, bien que l'âge médian était relativement jeune, ni en ce qui concerne leurs aptitudes auditives, qui n'étaient pas nécessairement développées par une activité musicale. L'expérience débutait par une émission sonore de dix secondes de bruit blanc à travers le casque stéréophonique de l'utilisateur. Ce bruit blanc avait pour but de masquer les bruits potentiels émis par la personne cible en mouvement pour rejoindre une position déterminée entre chaque transition. Après cette période de bruit blanc, les participants avaient la possibilité de pivoter sur leur chaise pour afin de déterminer la nouvelle position de la personne cible uniquement sur la base de

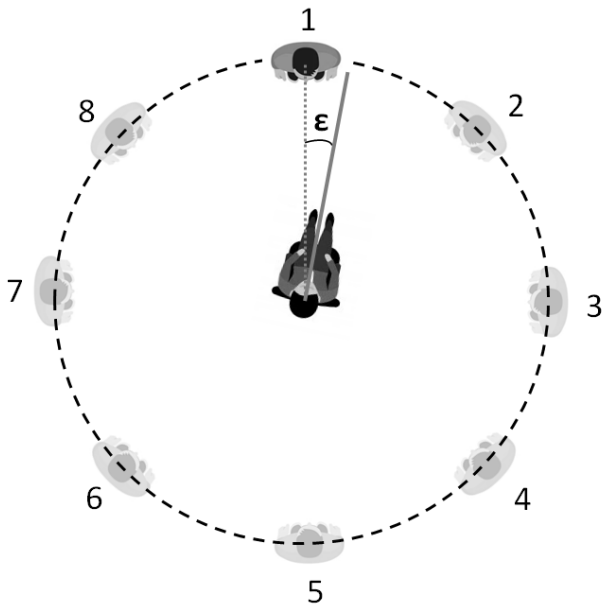


FIGURE 3.5 – Illustration du dispositif utilisé pour mesurer les capacités de localisation. Un participant est assis au centre de la zone d'expérimentation sur une chaise pivotante. Une personne debout change aléatoirement sa position parmi 8 emplacements marqués [1...8] espacés équitablement sur un cercle de deux mètres de rayon. L'erreur de l'angle azimutal ε est évaluée.

repères auditifs sans limite de temps. Après avoir réalisé cette tâche sur l'ensemble des huit positions, en se fiant à l'audition, les participants ont procédé à la même opération sans les yeux bandés, employant cette fois-ci la vision pour localiser la personne cible.

Les résultats présentés dans la figure 3.6 montrent la précision angulaire moyenne pour l'ensemble des huit positions cibles dans la condition auditive, c'est-à-dire avec le système sonore expérimental et les yeux bandés. Dans cette condition, l'erreur angulaire moyenne de localisation azimutale est de 6,72° ± 5,82 tandis que l'erreur angulaire moyenne, dans la condition, visuelle est de 2,85° ± 1,99 [152]. Une erreur angulaire deux fois supérieure a donc été mesurée dans la condition sonore. Cette augmentation était attendue compte tenu de la prédominance chez l'Homme de la modalité visuelle dans le traitement des informations spatiales. Néanmoins, la capacité des participants à localiser avec précision une personne en se basant sur le dispositif de substitution sensorielle auditive est significative. Cette faible erreur de localisation démontre une potentialité quant à l'utilisation de ce système pour transmettre d'autres informations spatiales pertinentes telles que la présence d'un obstacle ou une orientation à suivre pour atteindre une destination désirée.

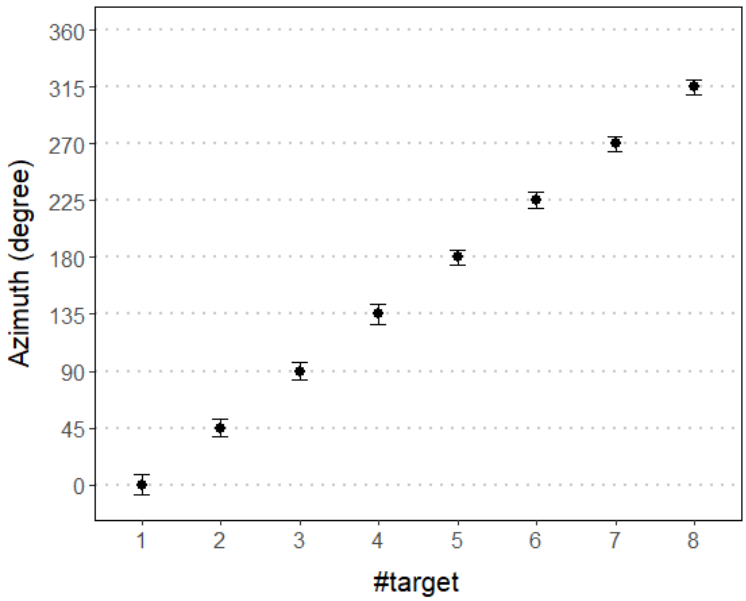


FIGURE 3.6 – Erreur moyenne associée de localisation pour chaque position azimutale avec un encodage de substitution sensoriel [152].

3.2.2 Performance de localisation de l'élévation

Comme précédemment mentionné, la précision de la position verticale à l'aide d'un son stéréophonique est moins bonne que pour la position horizontale. La combinaison d'indices acoustiques permet de faciliter l'interprétation du signal sonore. Similairement, à l'axe azimutal, nous avons réalisé une évaluation de l'acuité humaine de localisation d'une position verticale entre plusieurs méthodes d'encodages. Cette étude proposait une comparaison entre trois signaux sonore encodant l'information spatiale sur l'axe vertical. L'intérêt principal de l'étude réside dans l'évaluation de l'utilité de combiner deux indices acoustiques différents afin de favoriser une meilleure précision de localisation sur le positionnement d'un élément. Les trois signaux sonores utilisés pour encoder une information verticale ont les caractéristiques suivantes :

- **Signal sonore monotonique spatialisé :** Signal sonore décrit précédemment, dont l'élévation d'une information est fournie par une combinaison d'un indice spatial et fréquentiel pur.
- **Bruit blanc spatialisé :** Bruit blanc, d'une durée équivalente, dont l'indice acoustique lié à l'élévation est exclusivement fourni par une spatialisation par une paire d'HRTF 2D.
- **Signal sonore spatialisé avec une harmonique :** Version enrichie du signal mono-

tonique avec l'ajout d'une harmonique dans l'indice fréquentielle relatif à l'élévation.

Dans cette étude, la précision des encodages sonores pour la localisation verticale a été évaluée à l'aide de deux groupes distincts, de 18 participants chacun. Ces groupes étaient nommés "Monotonique" et "Harmonique". Le groupe Monotonique s'est concentré sur l'évaluation de l'élévation par une spatialisation d'une information avec un bruit blanc et par un encodage de l'axe vertical par la modulation de la fréquence d'un signal sonore monotonique. Le groupe Harmonique a réalisé la même tâche, mais voyait le signal monotonique enrichi par deux composantes fréquentielles. Pour chaque type d'encodage (bruit blanc, monotonique, harmonique), une évaluation de la précision a été effectuée au cours de 50 essais, avant et après une phase de familiarisation libre de 60 secondes. Durant ces essais, les participants devaient localiser et valider l'orientation d'un son, qui pouvait varier entre -25° et 25° avec des intervalles de 12.5°, en utilisant un pistolet en plastique sur lequel était placé un dispositif de suivi de position.

	Avant familiarisation	Après familiarisation
Bruit blanc (Groupe Monotonique)	40.19 ± 37.03°	24.90 ± 18.40°
Bruit blanc (Groupe Harmonique)	35.51 ± 33.69°	25.35 ± 20.31°
Signal monotonique	31.54 ± 27.19°	19.75 ± 16.25°
Signal harmonique	34.14 ± 33.69°	21.70 ± 16.72°

TABLE 3.1 – Erreur moyenne et écart type de la précision de localisation en élévation (en °) avant et après une session de familiarisation [153].

Le tableau 3.1 reporte la précision de localisation en élévation pour les différentes modalités d'encodages avant et après une session de familiarisation [153]. Ces résultats démontrent que l'exposition et la familiarisation avec l'ensemble de ces signaux améliorent significativement la capacité des participants à localiser précisément les sons en élévation, quel que soit l'encodage sonore employé. Néanmoins, il existe des variations de précision spécifiques à chaque méthode d'encodage. Les résultats montrent que la localisation des stimulus basés sur le bruit blanc, que ce soit dans le groupe Monophonique ou Harmonique, demeure relativement imprécise, même après familiarisation. L'introduction d'un indice fréquentiel en complément de l'information spatialisée réduit l'erreur dans l'interprétation de la position verticale. Cependant, l'ajout d'une composante harmonique pour rendre le signal plus riche n'apporte pas de progrès dans la capacité de localisation, voire donne des résultats plus contrastés que pour un signal monotonique.

3.3 Représentation des éléments proches

La proximité des éléments matériels d'une scène visuelle représentent une caractéristique majeure pour estimer leurs niveaux de dangerosité. Les systèmes de substitution sensorielle cherchent à offrir la perception rapide de la position de l'ensemble des objets proximaux. Pour cela, les signaux sonores sont générés à partir d'une carte de profondeur indiquant la distance relative entre les éléments de l'espace de navigation et la personne malvoyant. Plus précisément, la position de chaque pixel au sein de l'image renseigne sur l'azimut et l'élévation du détail filmé alors que son niveau de gris renseigne sur son éloignement. Cette image de profondeur est associée avec un seuillage des niveaux de gris correspondant à une distance *D* afin de filtrer et de retenir uniquement les objets situés à une distance inférieure ou égale à *D*. Les éléments se trouvant au-delà de cette distance spécifique sont représentés par des zones de pixels dont la teinte de gris est égale à zéro. Ces derniers sont désignés sous le terme de *pixels inactifs*. À l'opposé, les pixels qui sont retenus, appelés *pixels actifs*, subissent une normalisation de leurs valeurs, les faisant varier entre 0 pour une distance D et, 255 pour une distance très proche. Cette normalisation est effectuée en fonction de leur distance relative, comme le dépeint la courbe présentée dans la figure 3.7. Les pixels actifs sont ensuite traités par un algorithme d'encodage acoustique, s'inspirant de la méthode LibreAudioView [154] pour générer un signal sonore.

Le processus d'encodage de l'ensemble des pixels actifs en pixel sonores suit le diagramme de la figure 3.8. Ce processus génère un ensemble de pixels sonores qui sont ensuite additionnés tous ensemble pour constituer une *trame sonore globale*, par une

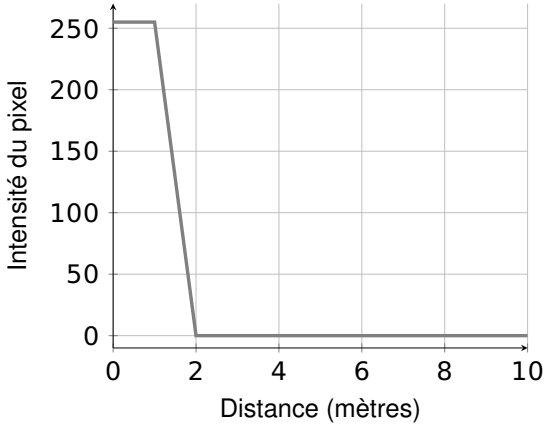


FIGURE 3.7 – Évolution de la nuance de gris du pixel visuel après une transformation en fonction de la distance relative à l'utilisateur.

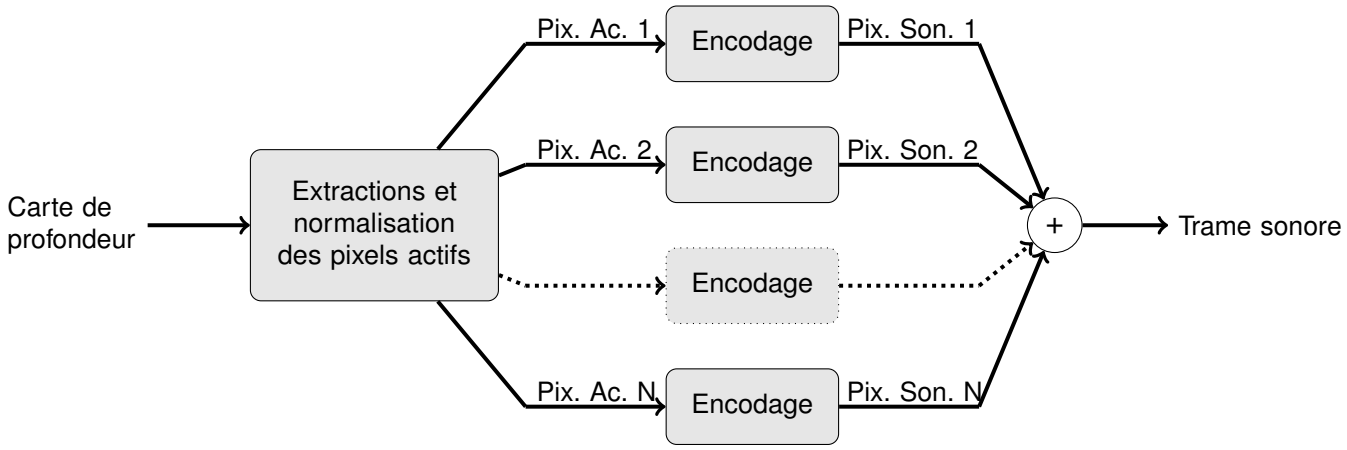


FIGURE 3.8 – Diagramme de la méthode d'encodage de l'ensemble des éléments proches en un unique signal sonore stéréophonique. Les termes *pix. Ac.* et *pix. Son.* correspondent respectivement à *pixel actif* et *pixel sonore*.

somme des différents signaux sonores. Cette méthode assure la conservation dans la trame globale des fade-in et fade-out (fondu en entrée et le fondu en sortie) présents en début et fin de chacun des pixels sonores. La proximité spatiale de l'objet par rapport à l'individu est traduite par une modulation de l'intensité du signal acoustique basée sur la teinte de gris du pixel associé. Dans cette configuration, un objet proche engendre une intensité sonore plus élevée, rendant l'information sonore plus saillante, tandis qu'un objet éloigné est traduit par une intensité sonore atténuée. Cette modulation est mise en œuvre

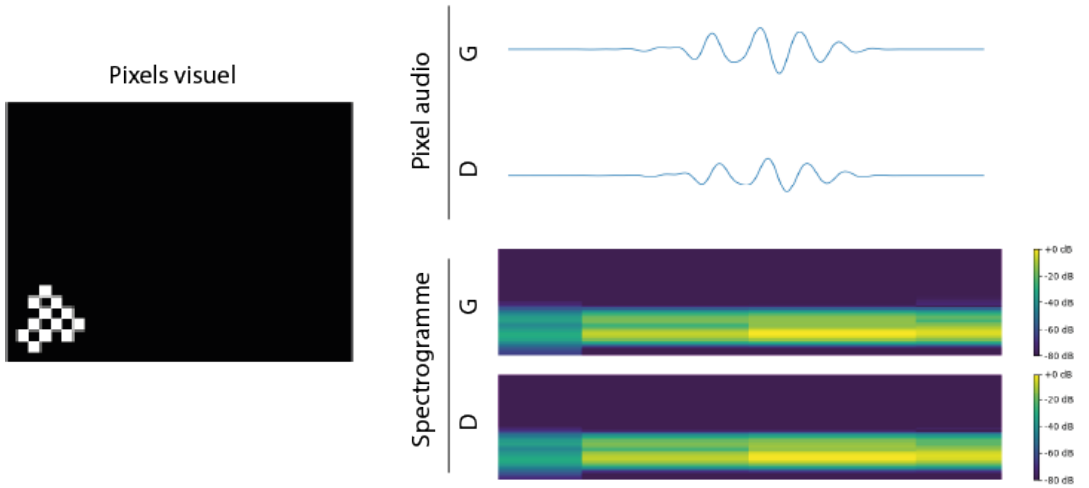


FIGURE 3.9 – Représentation d'une information visuelle située dans le coin inférieur gauche de l'image et de son interprétation sonore en termes de signal et de fréquence. Le signal sonore stéréophonique résultant de l'encodage des pixels actifs est d'une fréquence faible avec une amplitude du signal (position basse) du canal gauche plus élevée que le droit (élément à gauche de l'objectif).

grâce à une courbe de normalisation spécifiquement conçue pour ajuster l'intensité des pixels actifs. Cette courbe accorde une importance accrue lorsque l'objet est à proximité immédiate, et diminue progressivement cette intensité à mesure que l'objet s'éloigne. De plus, un mécanisme d'"alerte" est intégré : lorsque la distance d'un élément de la scène est inférieure à 1 mètre de l'utilisateur, la signature sonore est générée avec une intensité maximale, indiquant une proximité critique afin d'attirer immédiatement l'attention de l'utilisateur. Ce mécanisme permet d'assurer une réactivité accrue par la perception d'un signal sonore très saillant. Les figures 3.9 et 3.10 illustrent l'encodage sonore de deux informations visuelles distinctes. Les signaux stéréophoniques représentés montrent l'amplitude et la forme de l'onde, tandis que les spectrogrammes associés fournissent des informations sur la fréquence. Ces informations permettent d'estimer la configuration spatiale des éléments visuels en se fiant uniquement aux caractéristiques sonores.

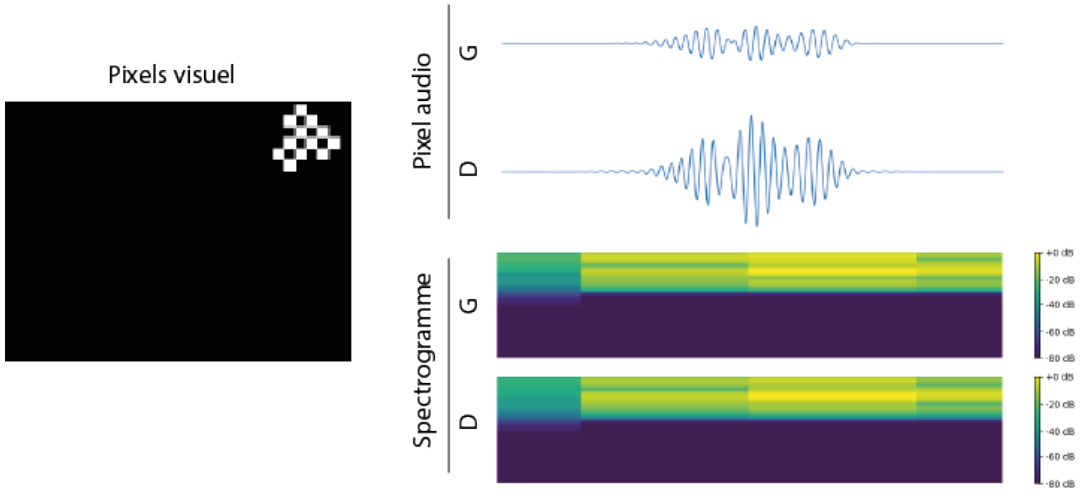


FIGURE 3.10 – Représentation d'une information visuelle située dans le coin supérieur droit de l'image et de son interprétation sonore en termes de signal et de fréquence. Le signal sonore stéréophonique résultant de l'encodage des pixels actifs est d'une fréquence élevée avec une amplitude du signal (Position haute) du canal gauche plus élevée que le droit (Élément à droite de l'objectif).

3.4 Preuve de concept

L'association d'une information sur les éléments environnants et sur la trajectoire à suivre est un objectif recherché pour de nombreux systèmes d'assistance à la mobilité pour les personnes aveugles. En effet, ces deux informations sont primordiales pour atteindre une destination particulière en toute sécurité. Cependant, la transmission des

informations relatives à ces deux aspects de manière séquentielle ou disjointe n'est pas évidente. La diffusion temporellement distincte de ces deux informations peut augmenter le temps nécessaire pour transmettre l'intégralité des indications essentielles et accentuer le caractère discontinu des informations sonores à traiter. De plus, un délai important dans l'émission accentue les hésitations d'une personne aveugle lorsqu'elle se déplace dans son environnement. Le stimulus sonore perçu peut ne plus être en phase avec l'orientation actuelle de la personne et donc perturber son analyse. La fusion de ces informations au sein d'une même trame sonore est ainsi rendue préférable afin de faciliter la prise de décision et évite des temps d'arrêt et d'ajustement. Dans cette optique, nous avons développé une preuve de concept dans le cadre d'une méthode et d'un système d'assistance à la navigation avec une faible latence pour permettre un traitement en temps-réel. L'objectif principal de ce système est d'évaluer la pertinence et l'efficacité de la fusion de ces informations sonores lors de la réalisation d'un itinéraire donné, en particulier dans un environnement intérieur.

3.4.1 Méthode de navigation dans un bâtiment

La navigation dans des espaces intérieurs est une tâche difficile pour les personnes aveugles, surtout en l'absence de connaissance préalable de l'agencement des locaux. Ces espaces, souvent constitués de nombreuses pièces aux configurations variées, sont occupés par de nombreux objets qui rendent l'environnement potentiellement dangereux. Néanmoins, les environnements intérieurs, par leur nature confinée, offrent une plus grande flexibilité et de plus larges possibilités de contrôle. Leur agencement peut être adapté pour répondre aux besoins spécifiques des personnes aveugles afin de rendre leurs déplacements plus intuitifs et sécurisés en incorporant, par exemple, des repères tactiles ou sonores. De plus, les caractéristiques spécifiques des espaces intérieurs peuvent être utilisées pour élaborer des méthodes d'assistance à la navigation combinant des techniques d'estimation de la trajectoire et de localisation de l'utilisateur. Parmi celles-ci, l'utilisation de balises sans fil telles que les RFID (*Radio Frequency IDentification*) [155], [156], le Bluetooth [67], [114], ou l'UWB (*Ultra Wide Band*) [157] est répandue. Ces dispositifs étant positionnés de façon à tracer un parcours de balise en balise jusqu'à la destination finale. Néanmoins, la mise en place d'un grand nombre de balises, nécessaire pour couvrir un espace complet, est souvent limitée par des coûts élevés, tant en termes de déploiement que de maintenance. Une alternative pour pallier ces contraintes repose sur l'utilisation d'informations visuelles, avec l'intégration d'éléments peu coûteux, facilement

détectables et modifiables selon les changements dans l'agencement de l'espace de navigation.

FIGURE 3.11 – Illustration d'un système de navigation sonore où l'utilisateur aveugle est guidé vers la position d'un marqueur visuel par l'émission d'un stimulus sonore.

Par conséquent, le système que nous avons développé pour aider les personnes malvoyantes à s'orienter et à se déplacer dans un bâtiment repose sur le placement stratégique de marqueurs visuels. Ce système présente des avantages significatifs, notamment sa simplicité et son coût réduit. Néanmoins, le positionnement de ces marqueurs nécessite une attention particulière. En les plaçant sur des éléments clés tels que les portes, ils peuvent indiquer des points de passage essentiels, facilitant ainsi la navigation à l'intérieur du bâtiment. De plus, une occlusion des marqueurs compromettrait la localisation et rendrait le système inefficace. Le maillage du bâtiment avec des marqueurs permet une schématisation de l'espace de navigation sous forme de graphe. Cette représentation facilite l'application d'algorithmes de recherche de chemin basés sur les graphes, mais nécessite une identification unique de l'ensemble des marqueurs visuels. Le fonctionnement global de la méthode de navigation du système de navigation en intérieur est décrit dans le diagramme 3.12 et se détaille comme suit :

- **Initialisation :** L'algorithme de recherche de chemin nécessite la connaissance du nœud de départ et du nœud d'arrivée pour s'initialiser. Le choix du nœud de départ est effectué en recherchant un marqueur visuel dans l'espace proche de l'utilisateur. Après avoir détecté le nœud de départ, l'utilisateur est invité à indiquer la destination

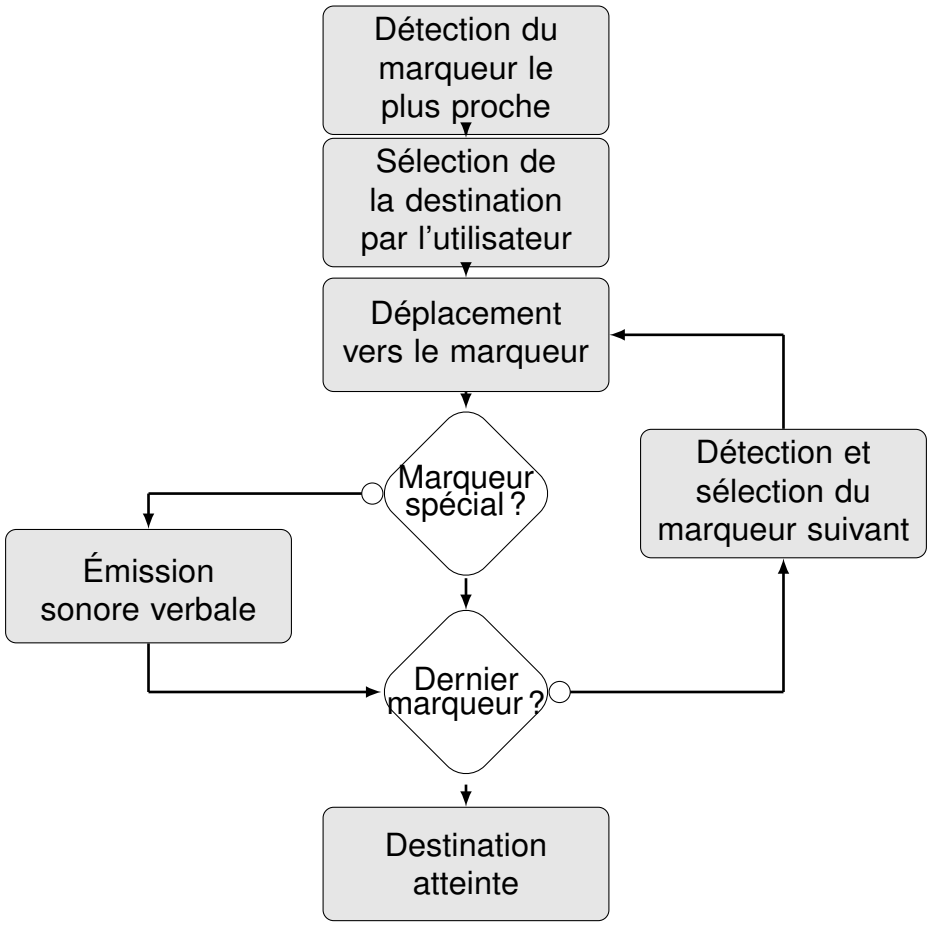


FIGURE 3.12 – Diagramme de l'algorithme de navigation intérieur.

souhaitée.

- **Déplacement de l'utilisateur :** L'utilisateur se déplace vers le marqueur sélectionné en suivant un son spatialisé symbolisant la position dans un espace sonore du marqueur visuel.
- **Atteinte du marqueur :** Lorsqu'un marqueur est suffisamment proche de l'utilisateur, le système vérifie si le marqueur est référencé comme un point d'intérêt à atteindre ou bien si une action de l'utilisateur est requise, comme ouvrir une porte. Ensuite, si le marqueur atteint n'est pas la destination finale, l'utilisateur est invité à scanner à nouveau l'environnement pour trouver le prochain marqueur.

Une erreur de détections de marqueurs peut empêcher l'utilisateur de poursuivre son chemin, tandis qu'un traitement lent des informations visuelles peut rendre la navigation moins fluide ou saccadée. Il est donc essentiel de choisir un marqueur doté d'un symbole unique et identifiable, ainsi qu'une méthode de détection associée qui soit robuste et efficace. Dans le cadre de cette preuve de concept, des marqueurs visuels nommés

STag [158] ont été utilisés. En effet, ces marqueurs ont été conçus pour être plus rapides, stables et efficaces dans la détection de marqueurs distants, même en cas d'occlusion partielle ou dans des conditions difficiles avec un angle de vue important. La détection des marqueurs STag repose sur l'extraction de caractéristiques géométriques simple telles que les ellipses, les coins et les bords, combinée à un raffinement homographique. L'identification des marqueurs Stag respecte ainsi une contrainte fondamentale des systèmes d'assistances avec un traitement temps réel. De plus, la vaste gamme de marqueurs individuels disponibles dans les bibliothèques permet une mise en œuvre étendue dans de grands espaces intérieurs. Un exemple de trois STags distincts est présenté dans la figure 3.13. D'autres méthodologies de balise visuelle, par exemple AprilTag [159] ou RUNE-Tag [160] existent, mais sont plus sensibles à une éventuelle occlusion partielle ou une distance de perception importante due à la finesse des motifs visuels. Cette dernière contrainte pouvant néanmoins être compensée par l'usage de marqueur de plus grande taille.

FIGURE 3.13 – Exemple de différents marqueurs visuel STag [158].

Le maillage du bâtiment, représenté sous forme de graphe non orienté, est constitué de nœuds interconnectés sans indication des distances les séparant. Cette structure de graphe non pondéré permet l'application d'algorithmes de recherche de chemin tels que la Recherche en Profondeur (Depth First Search, DFS) et la Recherche en Largeur (Breadth-First Search, BFS) pour identifier le chemin vers la destination. Ces algorithmes, bien que similaires dans leur objectif de trouver la destination, diffèrent dans leur méthode de parcours. Le DFS explore en profondeur, plongeant le plus loin possible dans chaque branche avant de revenir en arrière, tandis que le BFS examine le graphe en largeur, explorant tous les voisins d'un nœud avant de passer au niveau suivant. Bien que le DFS puisse être plus rapide et moins gourmand en mémoire, il ne garantit pas toujours le chemin le plus court, contrairement au BFS. Dans le cadre d'un système d'assistance aux personnes malvoyantes, l'algorithme BFS semble donc à privilégier. En effet, cette méthode assure que le chemin proposé est non seulement praticable, mais aussi le plus direct possible. Cette caractéristique est essentielle pour une navigation efficace et sûre

dans les environnements intérieurs, où il est important d'éviter des parcours inutilement longs et complexes, pouvant être fatigants et dangereux pour une personne aveugle.

Cependant, il est important de noter qu'une navigation linéaire simple n'est pas toujours réalisable dans des environnements complexes comportant de multiples obstacles, qu'ils soient statiques ou mobiles. Les espaces intérieurs, avec leurs divers obstacles et configurations, exigent souvent des approches de navigation plus sophistiquées, adaptées à la complexité et à la dynamique de ces environnements. En conséquence, bien que le BFS permette de trouver le chemin le plus court, le parcours indiqué doit pouvoir s'adapter à l'environnement pour naviguer efficacement et en sécurité.

3.4.2 Représentation auditive des données spatiales.

L'apport d'information annexe sur l'environnement de l'utilisateur lors de son trajet offrent une connaissance élargie de la configuration spatiale. L'unique indication de l'orientation à suivre n'est pas suffisante et peut isoler la personne malvoyante de son environnement. En effet, une signalisation des obstacles à proximité est nécessaire pour favoriser une meilleure compréhension de l'espace et un sentiment de sécurité accrue lors de déplacement. Cependant, la combinaison de deux indications sonores différentes peut perturber l'interprétation des signaux et réduire la confiance de l'utilisateur. La clarté et la distinction entre les signaux sonores représentant la trajectoire à suivre et ceux indiquant les obstacles sont essentiels pour le fonctionnement optimal du système. De plus, il est impératif que ces informations soient transmises simultanément et de manière continue, sans interruption ou alternance, afin de garantir une navigation sécurisée et fluide. La fusion de ces deux informations au sein d'un unique signal répond à ces contraintes. Les éléments proches représentant des obstacles sont extraits à partir d'une carte de profondeur puis encodés suivant la méthode d'encodage des éléments proches évoquée précédemment (section 3.3). Simultanément, la position du marqueur visuel à atteindre est traduite en un autre signal sonore qui encode sa position azimutale, agissant comme une boussole auditive pour l'utilisateur. La distinction entre ce signal et celui relatif aux obstacles est réalisée par la spatialisation d'un signal monophonique doté de caractéristiques acoustiques différentes. Ainsi, le signal $s_{obstacle}(t)$ relatif aux obstacles et défini dans l'équation 3.6 avertissant l'utilisateur d'un ensemble d'élément proche sera caractérisé par un signal monophonique basé sur un bruit blanc. Le signal indiquant la direction à suivre, $s_{orientation}(t)$, sera spatialisé à partir d'un son monophonique similaire à un bip aigu, partageant des similitudes sonores avec les signaux acoustiques employées pour les radars

de reculs. Ce signal se distingue par une tonalité plus élevée et une richesse spectrale plus large, ce qui le rend facilement différentiable des signaux utilisés pour signaler les obstacles. Le signal final combinant des deux signaux est formulé mathématiquement dans l'équation 3.6.

$$s_{global}(t) = s_{obstacle}(t) + s_{orientation}(t) \tag{3.6}$$

La figure 3.14 fournit une synthèse visuelle de la méthode d'intégration de ces deux signaux. Elle illustre, d'une part, la détection des obstacles (représentée sur la partie gauche de la figure) et, d'autre part, la direction à suivre (sur la partie droite). Les

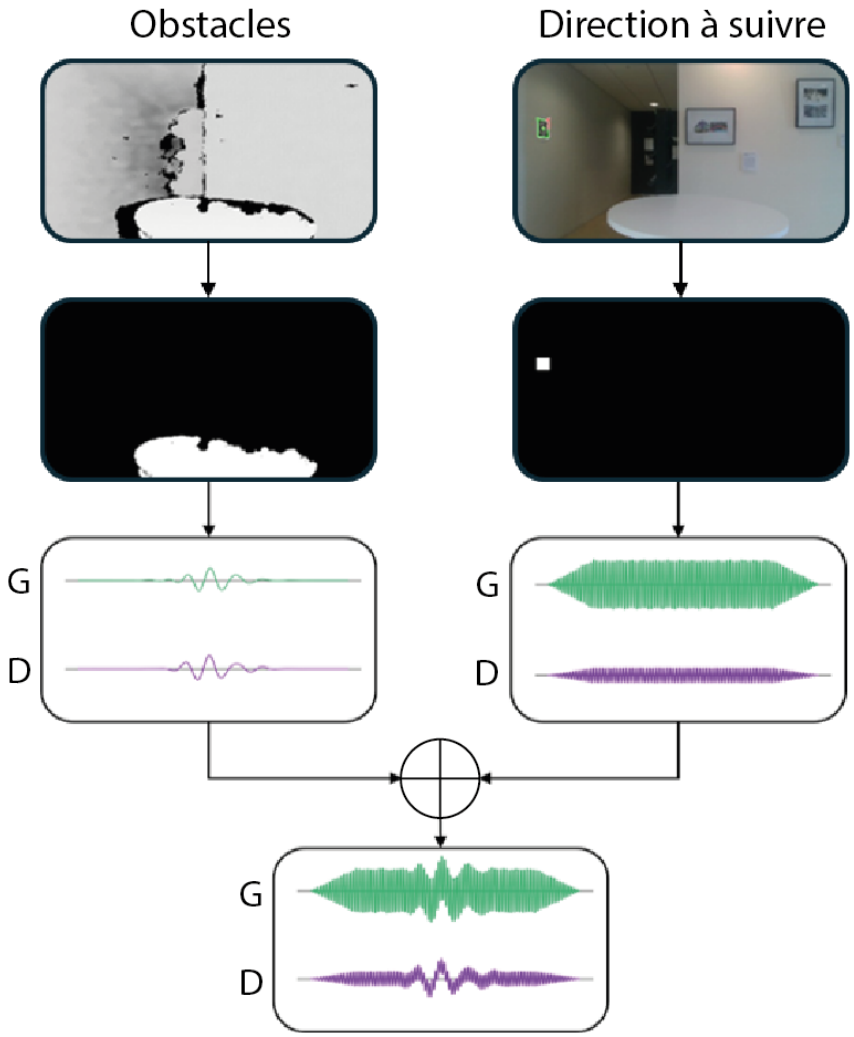


FIGURE 3.14 – Traitement vidéo et chaîne de sonification pour la détection d'obstacles (colonne de gauche) et la détection de marqueurs (colonne de droite). Les couleurs verte et violette symbolisent respectivement les canaux stéréophoniques gauche et droit.

informations sonores relatives aux obstacles sont spatialisées suivant la méthode définie précédemment (Chapitre 3.3) parallèlement à l'encodage de la position du marqueur à atteindre. Ces informations sont ensuite fusionnées en un unique signal sonore. Les courbes vertes et violettes symbolisent respectivement les canaux stéréophoniques audio gauche et droit. L'image inférieure montre la combinaison de ces sons, représentant le stimulus sonore transmis à l'utilisateur. L'image relative aux obstacles à proximité est définie à partir de la carte de profondeur afin d'obtenir les pixels actifs. Puis l'ensemble de ces pixels sont convertis en pixels sonores suivant leurs positions dans l'espace et leur distance par rapport à l'utilisateur (Seuil de perception à partir de 2 mètres, figure 3.7).

La présence d'obstacle dans la partie inférieure et centrale de l'image est représentée par un signal de faible fréquence et relativement similaire sur les deux canaux sonores. De même, la différence d'amplitude entre les canaux gauche et droit signaux de navigation indique la présence d'un repère à gauche de l'utilisateur. Enfin, ces deux signaux sont fusionnés en un seul signal global conservant l'ensemble des indices acoustiques relatifs aux informations de navigation et de détection d'obstacles.

En complément des sons spatialisés, des informations verbales non spatialisées sont fournies à des moments opportuns tout au long de la navigation. L'intégration de ces messages vocaux avec les signaux sonores spatialisés vise à pallier les limitations sémantiques inhérentes aux signaux sonores seuls. Ces messages verbaux fournissent à une personne malvoyante une description détaillée de son environnement de manière à le rendre facilement compréhensible pour accomplir une action donnée. Cette approche permet à l'utilisateur de mieux appréhender son parcours, en lui offrant une indication sur l'état d'avancement de son déplacement. De plus, elle lui permet d'être informé de la présence éventuelle d'une porte à ouvrir pour la poursuite de son trajet ou d'indiquer la présence d'un ascenseur, contribuant ainsi à une navigation plus fluide et sécurisée. Les identifiants de marqueur visuels sont exploités pour que l'utilisateur se situe devant un élément pertinent nécessitant une action. Ces informations verbales sont diffusées à des moments stratégiques de la navigation, que ce soit au commencement ou à la fin d'une étape, lorsque la personne est immobile et que le risque de collision avec un obstacle est relativement faible.

3.4.3 Expérimentation

Nous avons développé et conçu un système basé sur la méthode de localisation par marqueur et d'encodage sonore pour évaluer le comportement d'une personne dépourvue

de connaissance visuelle dans un bâtiment assistée par notre méthode. Ce système avait pour objectif principal d'évaluer les capacités humaines de localisation et d'interprétation de l'ensemble des informations sonores émises afin de définir la viabilité d'un tel encodage. L'expérience que nous avions conduit impliquait de guider un participant les yeux bandés, simulant ainsi un handicap visuel, à vers une destination inconnue. Le principal défi résidait dans la capacité du participant à se repérer et à se déplacer en se fiant uniquement au système d'assistance, sans recours à des aides visuelles ou physiques. Un superviseur l'accompagnait à une certaine distance pour l'observer et l'assister en cas de problème afin de garantir des conditions expérimentales sécurisées. Avant de commencer, une explication concise du fonctionnement du système et des caractéristiques des signaux sonores émis ont été fournis au participant. Dans ce contexte, un balisage d'un environnement réel et existant en conservant sa structure d'origine a été effectué pour obtenir un environnement de navigation réaliste.

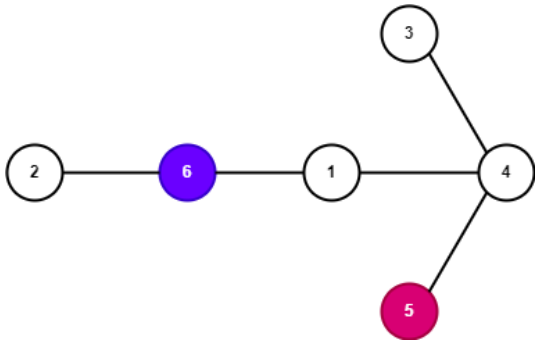


FIGURE 3.15 – Représentation schématique sous forme de graphe de l'interconnexion des marqueurs visuels dans l'espace de navigation. Le nœud violet indique la présence d'une porte symbolisée par un marqueur spécial.

Cet environnement a été choisi pour reproduire les conditions de navigation quotidiennes des personnes malvoyantes. Le lieu de l'expérience était le 3^e étage d'un immeuble de bureaux (Bâtiment I3M, occupé par les laboratoires ImViA & LEAD), à Dijon. Cet étage comportait divers obstacles statiques, tels que des chaises, des bureaux et des tables, créant un parcours jonché d'obstacles pour le participant. Le balisage a été réalisé avec l'installation de six marqueurs imprimés sur papier A4, placés à des endroits stratégiques. Ces marqueurs formaient un réseau de repères, essentiel pour le fonctionnement algorithmique du système de navigation. Une partie du maillage de l'étage du bâtiment est représenté sous une forme schématique dans la figure 1, illustrant en détail le maillage et son utilisation dans l'algorithme de recherche de chemin. Dans cette représentation fournissant une référence visuelle pour comprendre les interconnexions à travers le bâtiment, la présence d'une porte est indiquée par un nœud violet.

Au début de l'expérience, le participant était placé près du marqueur numéro 2, situé dans une pièce spécifique du bâtiment. L'objectif était de naviguer vers l'ascenseur de l'étage, symbolisant une sortie du bâtiment, indiqué par le marqueur numéro 5. Le chemin optimal, défini par l'algorithme de recherche de chemin, était tracé de manière à suivre la séquence des marqueurs $2 \rightarrow 6 \rightarrow 1 \rightarrow 4 \rightarrow 5$. Le marqueur numéro 3, positionné dans l'espace central, avait un rôle spécifique. Il était prévu comme un leurre, destiné à dérouter le participant vers une mauvaise destination en cas d'échec de l'identification des marqueurs. En naviguant sur ce parcours, le participant devait contourner divers obstacles. Parmi ceux-ci, le passage d'une porte, signalée par le marqueur n° 6, ainsi que plusieurs éléments de mobilier, positionnés entre les marqueurs n° 1 et n° 4, ainsi qu'entre n° 4 et n° 6.

Le parcours emprunté par l'utilisateur pour atteindre le marqueur 5 a été enregistré afin d'analyser la viabilité du système d'aide à la navigation en intérieur, et plus particulièrement pour observer le comportement de l'utilisateur face aux obstacles. La position relative de la personne dans le bâtiment a été suivie et enregistrée tout au long de son déplacement. Les variations observées dans la trajectoire de l'utilisateur sont attribuables à deux facteurs principaux. Premièrement, elles sont dues au balancement naturel du corps humain pendant la marche. Deuxièmement, elles résultent des mouvements de la tête de l'utilisateur, nécessaires pour balayer l'environnement à la recherche d'éventuels obstacles ou pour déterminer avec précision l'angle azimutale relatif à l'orientation du marqueur cible. En parallèle de notre système de navigation, un casque Oculus Quest 2 était porté par l'utilisateur afin d'enregistrer ses mouvements, par l'intermédiaire du système de tracking positionnel intégré. Le changement de marqueur pour passer au suivant est effectué lorsque l'utilisateur est à moins de 80 cm. Cette distance de 80 cm, le protège d'un contact avec le mur sur lequel se situe le marqueur, bien que celui-ci soit perçu comme un obstacle proche avec une identité sonore particulière. En effet, la détection sonore d'un mur est identifiable par la présence d'un son spatialisant un obstacle sur une large surface avec une amplitude sonore uniforme. Ce suivi a révélé des aspects intéressants du comportement de l'utilisateur, notamment la manière dont il se repositionne légèrement avant d'ouvrir une porte, ainsi que sa capacité à détecter et à éviter les tables en se dirigeant vers les marqueurs n°4 et n°5. La figure 3.16 fournit une vue aérienne du bâtiment, obtenue grâce à plusieurs scans Lidar. Cette représentation inclut des informations détaillées telles que les emplacements des marqueurs visuels, indiqués par des flèches de couleur verte et violette. Elle montre également la position de départ du participant, marquée par un point rouge, et la position d'arrivée, matérialisée en rose. Le chemin suivi par le participant est représenté par une ligne blanche, illustrant le trajet

parcouru depuis le point de départ jusqu'à la destination finale, tout en naviguant à travers les obstacles et en suivant les indications du système de navigation.

Les résultats de l'expérience démontrent que l'utilisateur, même privé de la vue, peut naviguer avec succès dans un environnement intérieur, en évitant les obstacles statiques (tels que les murs), et en franchissant des portes pour atteindre la destination souhaitée à partir d'un signal sonore associant une information sur la trajectoire à suivre et les obstacles à proximités. L'utilisateur a ressenti un réel confort durant son déplacement dans la compréhension de l'action à réaliser par l'émission des informations verbales personnalisées en fonction de l'identifiant du repère visuel. En effet, les détails utiles et précieux sur la caractéristique de la porte ont permis à l'utilisateur non seulement de naviguer physiquement dans l'espace, mais aussi de recevoir des informations contextuelles qui enrichissent son expérience de navigation. L'ajout de marqueurs dans un espace permet, outre l'assistance à la navigation, d'enrichir la connaissance de l'espace en associant un message verbal personnalisé à chaque élément clé de l'espace.

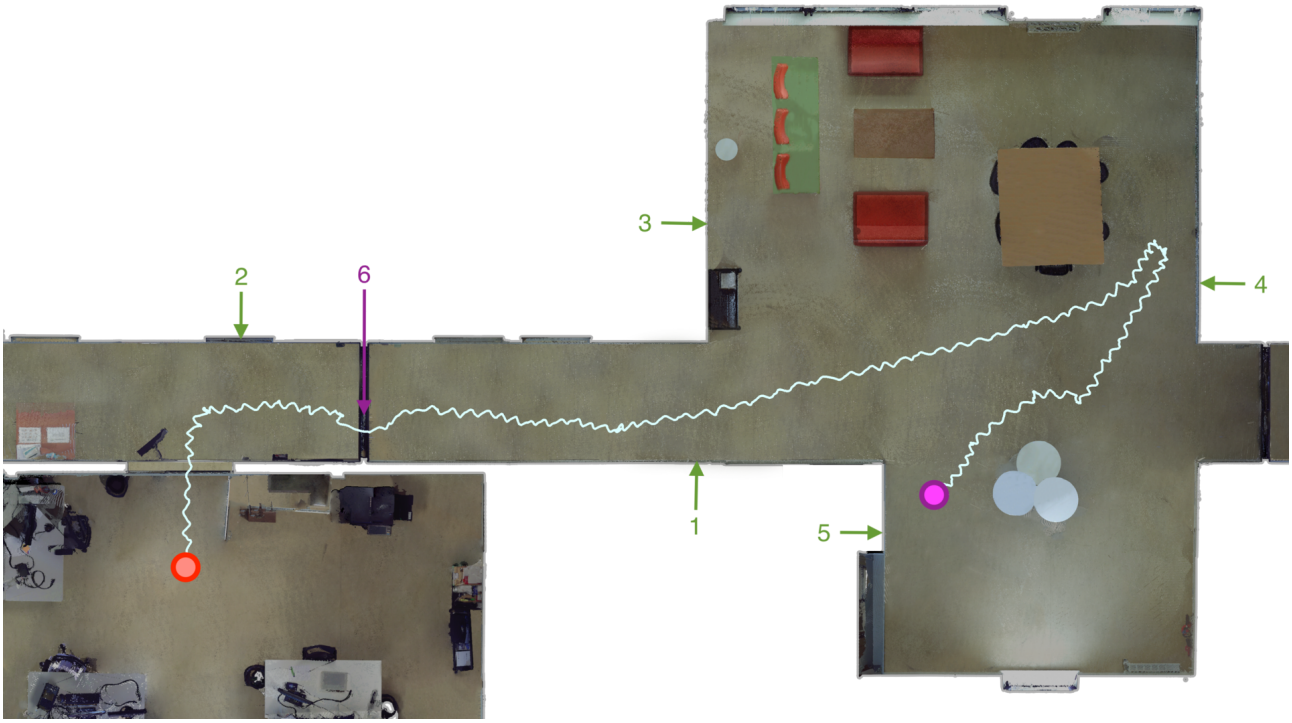


FIGURE 3.16 – Vue de dessus des mouvements de l'utilisateur pour atteindre la destination souhaitée. Le point rouge et les points roses indiquent respectivement la position de départ et la destination. Les flèches représentent la position des marqueurs dans l'espace du bâtiment et la flèche violette indique que le marqueur est placé sur une porte. Le numéro associé à une flèche représente le numéro d'identification du marqueur. La vue de l'environnement de navigation a été capturée a posteriori, en utilisant plusieurs prises de vue par LIDAR.

3.5 Conclusion

L'encodage d'informations visuo-spatiales dans le domaine auditif offre l'accès à des informations spatiales non perceptibles naturellement pour une personne malvoyante pour améliorer sa compréhension de son environnement et, par extension, la vie quotidienne. L'efficacité de cette méthode dépend largement de la façon dont l'encodage sonore des informations est adapté aux capacités humaines d'analyse des stimuli sonores pour localiser leur origine. Bien que l'entraînement à long terme puisse améliorer significativement la capacité de l'utilisateur à interpréter ces signaux, l'utilisation des indices acoustiques utilisées naturellement par le système auditif humain pour identifier l'origine d'un son joue un rôle clé dans l'efficacité de cette méthode. En particulier, il apparaît qu'une spatialisation stéréophonique et l'atténuation du son sont facilement interprétées grâce à leurs similarités avec la perception auditive naturelle.

L'usage de ces mécanismes d'encodages visuo-sonore permettent d'informer l'utilisateur sur des détails essentiels nécessaires à l'accomplissement d'actions autonomes comme une assistance à la navigation. En effet, la navigation dans une zone non familière est une tâche ardue, voire impossible pour une personne malvoyante devant distinguer à la fois la trajectoire à suivre et les obstacles à proximité. Cependant, la transmission auditive de ces deux informations pour compenser les limitations visuelles exige une réception simultanée et exacte par l'utilisateur afin de rendre la navigation fluide et sûre. Une méthode d'encodage de ces éléments en de courts signaux spatialisés pour la réalisation d'une tâche de localisation a été proposée et évaluée [152], [153]. Ces éléments ont été fusionnés en un même signal en les maintenant distinguables tout au long du déplacement. Cette approche a été validée à travers le développement d'un système d'assistance à la navigation en intérieur, conçu comme un prototype pour répondre aux défis rencontrés par les personnes aveugles dans cet environnement. Un test en environnement réel a permis d'évaluer l'efficacité de cette méthode, démontrant son potentiel significatif pour améliorer l'autonomie et la sécurité des personnes malvoyantes au quotidien [161]. La personne équipée du dispositif expérimental a réussi à atteindre une destination inconnue en évitant de manière sécurisée les obstacles sur son chemin.

L'utilisation de méthodes de substitution sensorielle, convertissant les informations visuelles en signaux sonores, s'avère prometteuse pour enrichir la perception des personnes malvoyantes, particulièrement en termes d'assistance à la navigation. L'exploitation accrue de l'information visuelle pour identifier des éléments importants via des méthodes de vision par ordinateur, comme la perception de dangers ou la détermination de zones accessibles

pour se déplacer, peut ouvrir la voie à une plus grande autonomie des personnes aveugles dans des milieux non familiers. En effet, des espaces plus complexes et dangereux à appréhender tels que les environnements extérieurs nécessitent une analyse plus approfondie de l'espace de navigation par une connaissance visuo-spatiale des éléments environnants de la personne malvoyante. L'intégration d'informations supplémentaires représente un pas significatif vers l'amélioration de l'indépendance et de la qualité de vie des personnes ayant des déficiences visuelles, en particulier pendant leurs déplacements à travers des espaces non familiers ou dangereux.

4

Analyse sémantique d'un espace visuel

Sommaire

La perception visuelle joue un rôle crucial pour l'homme dans l'interprétation de son environnement lors de mouvements ou d'actions. L'absence ou la diminution de cette capacité visuelle représente un obstacle quotidien, réduisant l'autonomie des personnes malvoyantes. La conversion d'informations visuelles, inaccessibles en raison d'un handicap visuel, en signaux sonores via un mécanisme de substitution sensorielle, offre un potentiel significatif d'amélioration de la qualité de vie de ces personnes, en particulier en ce qui concerne leur mobilité. Les méthodes de substitution sensorielle visuo-sonore ont démontré leur capacité à localiser un ou plusieurs éléments dans un espace, compensant ainsi en grande partie le manque d'informations visuelles. Cependant, ces informations sonores se limitent à une connaissance des éléments à proximité d'une personne malvoyante. L'enrichissement de ces informations perçues par une analyse des données acquises offre une ouverture vers une compréhension accrue de l'environnement comme la nature d'un obstacle ou l'identification de zones de déplacement. La connaissance amplifiée de leurs environnements confère une meilleure autonomie et une qualité de vie, facilitant une interaction et une mobilité plus aisée en toute autonomie et sécurité. De surcroit, cet aspect fondamental dans des milieux confinés et régulés comme les espaces intérieur, s'avère critique dans des contextes extérieurs plus complexes et potentiellement plus risqués à appréhender sans assistance extérieure.

Dans les environnements extérieurs, les personnes malvoyantes sont confrontées à des difficultés nombreuses et variées. La navigation dans de grands espaces ouverts peut s'avérer complexe, en particulier en l'absence de repères visuels évidents. Ces espaces comprennent souvent des zones à risque, comme les voies de circulation automobile et les passages piétons, qui constituent une source de stress significative. Contrairement aux environnements intérieurs, où les obstacles sont généralement statiques, en extérieur, les personnes malvoyantes doivent composer avec des éléments dynamiques et potentiellement dangereux, tels que les voitures ou les trottinettes en mouvement. Un autre aspect complexifiant ce milieu est la diversité des configurations urbaines et du mobilier urbain. Cette hétérogénéité peut rendre difficile la formation d'une compréhension cohérente de l'environnement, exacerbant ainsi les défis liés à l'orientation et à la sécurité. Cependant, ces freins sont réduits chez une personne voyante due à une capacité de perception visuo-spatiale définissant la nature et la localisation des éléments environnants vitaux. La substitution des mécanismes naturels par des méthodes de vision artificielle avancées et adaptées au contexte permet une compensation de ces difficultés inhérentes aux personnes malvoyantes dans des espaces non familiers et complexes.

Ce chapitre se consacre aux méthodes d'extraction et d'analyse d'informations visuelles destinées à faciliter le quotidien des personnes malvoyantes lors de leurs dépla-

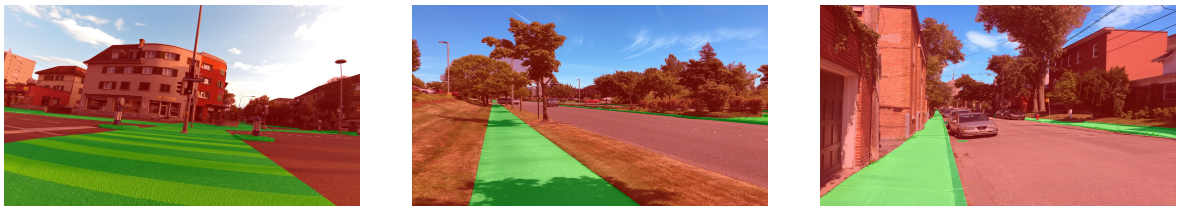


FIGURE 4.1 – Exemple de discrétisations des zones vertes et zones rouges sur des scènes urbaines.

cements en milieu urbain. Il explore les méthodes employées pour identifier et localiser des éléments clés dans cet environnement, tels que les éléments dangereux comme les obstacles physiques ou les zones non accessibles aux piétons. En effet, ces freins réduisent la capacité des personnes malvoyantes à se mouvoir en toute sécurité dans un espace et requièrent une grande attention pour reconnaitre la présence de zones sûres pour une navigation piétonne. Ces zones dépourvues d'obstacle sont appelées « zones vertes ». À l'inverse, la présence d'élément comme les poteaux, les arbres ou les routes pour véhicules constituent des dangers significatifs et doivent être appréhendée par le système ou la personne à l'aide d'alerte sonore. Ces zones à risque sont désignées comme "zones rouges". La figure 4.1 présente trois exemples de scènes urbaines, mettant en lumière ces zones vertes et rouges à travers l'usage de masques colorés.

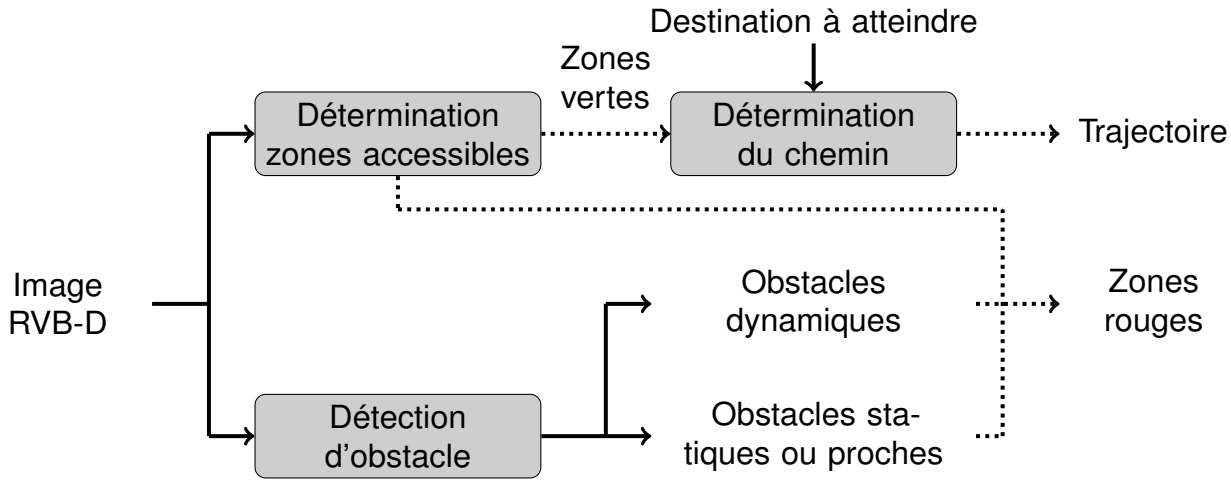


FIGURE 4.2 – Diagramme de traitement des données visuelles. L'image RVB-D est utilisée pour définir la présence d'obstacle et des zones accessibles aux piétons.

Cette séparation des zones de l'espace urbain en deux groupes, vert ou rouge, permet d'identifier les zones accessibles de l'espace vers lequel un déplacement est possible lors d'un suivi d'itinéraire pour atteindre une destination souhaitée afin d'éviter les zones incertaines. La figure 4.2, initialement présenté dans un cadre plus large à

la figure 2.25, dépeint le processus de détermination des zones vertes ou rouges pour une personne dans un milieu urbain non familier à partir d'informations visuelles. Le processus se déploie en deux branches parallèles et complémentaires, permettant une évaluation exhaustive de l'environnement : la première branche se focalise sur l'examen des éléments mobiles et statiques entourant la personne, alors que la seconde s'attelle à l'identification des zones vertes accessibles, cruciales lors d'un déplacement pour atteindre une destination souhaitée en toute sécurité. Ce chapitre traitera ces deux axes en débutant par une analyse de la localisation et de la dangerosité des éléments situés autour d'une personne malvoyante, pour ensuite se concentrer sur la détermination des zones accessibles aux piétons.

4.1 Localisation des éléments pertinents d'un espace urbain

La connaissance de la position d'un élément et de sa nature est cruciale lors d'un déplacement dans un espace urbain. Bien que la canne blanche puisse aider à identifier certains objets, elle ne fournit qu'une compréhension partielle de l'environnement, réduite à la portée et à la hauteur de la canne, pouvant être limitée lors de certaines circonstances ou contextes. Le franchissement d'une voie routière à un passage piéton ou l'évitement d'un obstacle potentiellement dangereux nécessite une connaissance plus étendue de l'environnement d'une personne. Malheureusement, ces éléments restent souvent non perçus par une personne malvoyante. De plus, ces émissions sonores tendent à se réduire avec l'introduction de moyen de transport plus silencieux tels que les trottinettes et les voitures électriques. Ces véhicules, bien que bénéfiques pour réduire la pollution sonore, diminuent la capacité des personnes aveugles à les percevoir alors que leur vitesse de déplacement représente une adversité supplémentaire pour les personnes ayant une déficience visuelle due à leurs vitesses de déplacement. En effet, les éléments dynamiques peuvent présenter un risque accru par rapport à des éléments statiques et doivent être perçus en amont et en priorité afin d'être correctement identifiés et contournés si besoin.

Les méthodes par apprentissage profond, en particulier les méthodes de détection d'objet évoquées dans la section 2.3.2.2 offrent des mécanismes de localisation d'objet présent sur une image avec précision illustrée à la figure 4.3. Ces méthodes retranscrivent la position bi-dimensionnelle d'un élément à partir d'un vecteur renseignant le point central de l'objet ainsi que sa hauteur et largeur symbolisées par une boite englobante. Dans le cadre d'une détection multi-étiquette, l'information de position est complétée par un attribut

FIGURE 4.3 – Illustration d'un résultat obtenue par une méthode de détection d'objet par apprentissage profond.

décrivant la nature ou une caractéristique spécifique de l'élément, permettant ainsi de le distinguer des autres. Cette information enrichie facilite une interprétation sémantique de la scène, essentielle pour identifier les éléments particulièrement dangereux pour l'utilisateur. De plus, la précision de ces méthodes peut être associée dans le cas de certaines architectures développées avec un temps de prédiction ou d'inférence faible requis dans le cas d'un système d'assistance à la mobilité pour personne malvoyante où le temps d'analyse et d'encodage sonore doit être optimisé et réduit au minimum pour offrir un temps d'interprétation sonore suffisant pour l'utilisateur. Cependant, les méthodes de détection d'objet par apprentissage profond sont principalement des techniques par apprentissage supervisé nécessitant des données annotées spécifiques au contexte d'utilisation.

4.1.1 Ensemble de données dans le cadre d'une navigation urbaine et piétonne

4.1.1.1 Jeux de données existants

Les jeux de données sont le socle des méthodes par apprentissage supervisé permettant à un modèle paramétrique d'ajuster ses paramètres afin de généraliser un phénomène à partir d'exemples spécifiques. Cette approche, amplifiée dans le cadre des méthodes par apprentissage profond, nécessite une large collection d'échantillons étiquetés et représentatifs pour entrainer efficacement le modèle à reconnaître et à prédire les caractéristiques pertinentes dans de nouvelles données. Dans le cadre d'une approche

FIGURE 4.4 – Image synthétique générée à partir d'une modèle 3D de la place Darcy à Dijon.

visant à localiser les obstacles présents dans des zones urbaines pour les personnes malvoyantes, les annotations doivent méticuleusement refléter les objets typiques et significatifs de ces milieux. Le tableau 2.2 présenté précédemment résume une sélection de jeux de données existants, incorporant des éléments annotées de la vie quotidienne aussi bien dans des environnements extérieurs qu'intérieurs. Cependant, les ensembles de donnée axés sur des exemples urbains sont principalement orientés pour le développement des véhicules autonomes, sans parfois intégrer des objets présents sur un espace piéton, à l'exception du jeu de données SideGuide [133]. Cet ensemble de données a été conçu pour soutenir le développement de dispositifs d'assistance pour les personnes malvoyantes, incluant l'annotation de 29 classes d'objets fréquemment rencontrés dans une ville de Corée du Sud. Néanmoins, des variations entre différents environnements urbains peuvent restreindre la capacité de généralisation du modèle, notamment en ce qui concerne la disposition des éléments ou leur conception, comme les panneaux de signalisation ou les feux pour piétons. De même, les conditions environnementales d'acquisition des données telle que la présence de soleil ou de pluie peuvent également altérer les performances.

4.1.1.2 Création d'un nouveau jeu de donnée : uB-Geoloc

En complément de cette base de donnée existante, nous avons créé un nouvel ensemble de données nommé uB-Geoloc [162] pour améliorer la robustesse des méthodes d'apprentissage face à différents contextes d'utilisation et favoriser le développement de technologies d'assistance aux personnes malvoyantes dans un milieu urbain. Cette nouvelle base est structurée autour de 16 séquences de données visuelles, collectées dans la ville de Dijon, et couvre une variété d'espaces de navigation, incluant des rues

étroites, de vastes zones piétonnes, et des jardins publics moins densément peuplés en éléments. Ces scènes dépeignent des situations courantes du quotidien en extérieur, telles que traverser un passage piéton, naviguer dans une foule ou marcher le long d'un trottoir. Bien que ces situations semblent simples, elles présentent des défis importants pour les personnes malvoyantes, en raison de la diversité et de la complexité des environnements.

Ces scènes ont été acquises à partir d'environnement réel et synthétique complémentaire. En effet, les données réelles visent à reproduire la vision d'une personne dans un environnement urbain avec une richesse et une authenticité propre aux données réelle. Ces séquences vidéos sont complétées par deux séquences synthétiques, générées via des simulations virtuelles. Ces données, générées par un moteur de jeu, fournissent une représentation contrôlée et uniformisée des obstacles, ainsi que des scénarios à risque difficiles à reproduire dans des conditions réelles. Néanmoins, une continuité entre ces deux types de données est réalisée par l'acquisition des données synthétiques à partir d'un modèle virtuel de la place Darcy à Dijon. Ce modèle 3D à faible polygone, basé sur des nuages de points acquis par scan LIDAR, a été enrichi par l'enregistrement des mouvements de tête réalisé avec un casque de réalité virtuelle (Oculus Quest 2.0) dans un gymnase vide. Le modèle 3D et les trajectoires de la tête ont ensuite été intégrés dans le logiciel 3DSMax pour un rendu graphique proche de l'aspect de la place.

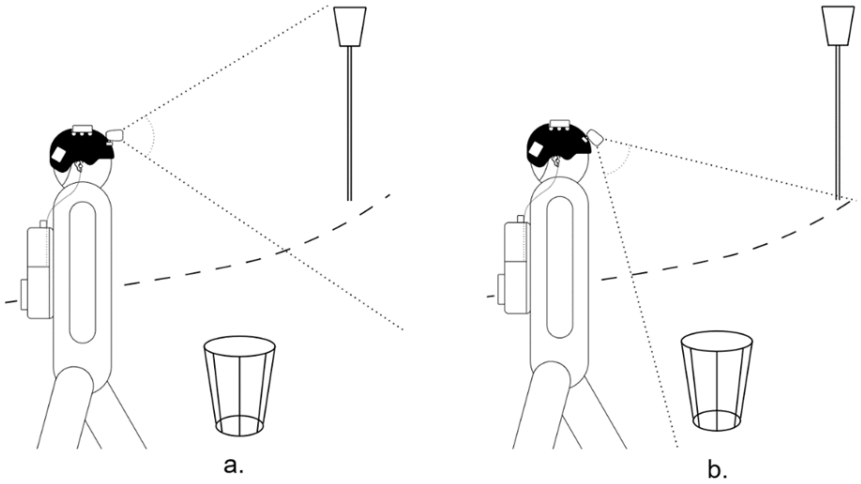


FIGURE 4.5 – Vue schématique illustrant diverses positions d'élévation de la caméra. La figure **a** correspond à un angle d'élévation de 0°, tandis que la figure **b** représente un angle d'élévation de -40°.

Scène	Nature	Élévation	Nombre d'images	Description de la scène
1	Synthétique	0°	1609	Franchissement d'une route dense à proximité d'une zone piétonne.
2	Synthétique	0°	900	Navigation dans une zone piétonne composée d'obstacles.
3	Réel	-40°	900	Déplacement sur un chemin piéton protégé.
4	Réel	-40°	900	Présence d'un escalier et d'obstacles exclusivement statiques.
5	Réel	-40°	900	Zone résidentielle avec une densité élevée d'obstacles statiques.
6	Réel	-40°	900	Zone résidentielle avec des obstacles statiques et dynamiques.
7	Réel	-40°	900	Avenue dense en véhicules et personnes statiques ou mobiles.
8	Réel	-40°	1597	Traverser d'un carrefour routier important avec feux et passages piétons.
9	Réel	-40°	900	Situation d'un bus s'arrêtant à un arrêt le long d'un trottoir.
10	Réel	-40°	804	Esplanade piétonne arborée avec un nombre d'obstacles faibles.
11	Réel	-40°	900	Large voie piétonne arborée avec une grande variété d'obstacles.
12	Réel	-40°	5522	Centre-ville urbain à proximité d'un parc et d'un lycée.
13	Réel	0°	6547	Multiples traversées de routes avec feux et passages piétons.
14	Réel	0°	8579	Navigation en présence de nombreux obstacles
15	Réel	0°	5899	Rue à sens unique exiguë et en travaux.
16	Réel	0°	9344	Place arborée avec plusieurs terrasses de bar.

TABLE 4.2 – Description de chaque séquence d'image identifiée par un numéro d'index. La colonne nature indique si la scène est réelle ou virtuelle. La colonne élévation fournit des détails sur la position angulaire de la caméra, avec 0° représentant une caméra parallèle au sol, et la colonne description de la scène résume chaque scène.

La perception des éléments par une caméra, qu'elle soit virtuelle ou réelle, dépend largement de son champ de vision. Contrairement à l'œil humain, qui peut diriger son regard vers différentes zones de l'espace, le champ de vision de la caméra est restreint aux mouvements de la tête sur laquelle est monté le dispositif d'enregistrement, réelle ou synthétique. Cette limitation est en partie compensée par une approche d'enregistrement des séquences vidéo sous différents angles d'élévation de la caméra comme illustré à la figure 4.5. Cette figure montre deux configurations d'orientation de la caméra et son impact sur les éléments observés. Un champ de vue plus large donne une indication globale sur la scène filmée tandis qu'une élévation angulaire plus basse priorise la perception des éléments proches dangereux pour l'utilisateur. L'ensemble des caractéristiques des différentes séquences enregistrées sont retranscrites dans le tableau 4.2 incluant l'angle de capture de la scène, mais également la nature, le nombre d'images et une brève description du contexte de navigation.

FIGURE 4.6 – Exemple de deux images issues du jeu de données annotées. Chaque classe est représentée par une couleur de boîte englobante différente.

Ces séquences de données, illustrant des scénarios courants lors de déplacements en milieu urbain, ont été soigneusement annotées pour faciliter leur utilisation par méthodes d'apprentissage supervisé, particulièrement pour la détection d'objets. Cette étape d'annotation s'était concentrée sur l'étiquetage de 28 éléments caractéristiques. Ces objets, décrits ci-dessous, correspondent aux éléments statiques ou dynamiques les plus significatifs en raison de leur fréquence dans les zones urbaines ou du risque qu'ils posent pour les personnes aveugles. Cette sélection assure que les modèles d'apprentissage puissent efficacement reconnaître et interpréter les éléments les plus communs et les plus menaçants de ces milieux. Un exemple d'annotation sur deux images issues de la base de donnée est illustré à la figure 4.6.

- **Liste des classes :** Arbre, Banc, Table, Chaise, Borne à incendie, Poubelle, Panneau signalétique, Feu tricolore, Poteau, Panneau mobile, Arrêt de bus, Feu piéton,

Passage piéton, Barrière, Support à vélo, Parasol et Plante, Voiture, Camion, Bus, Moto, Trottinette, Vélo, Personne et Animal.

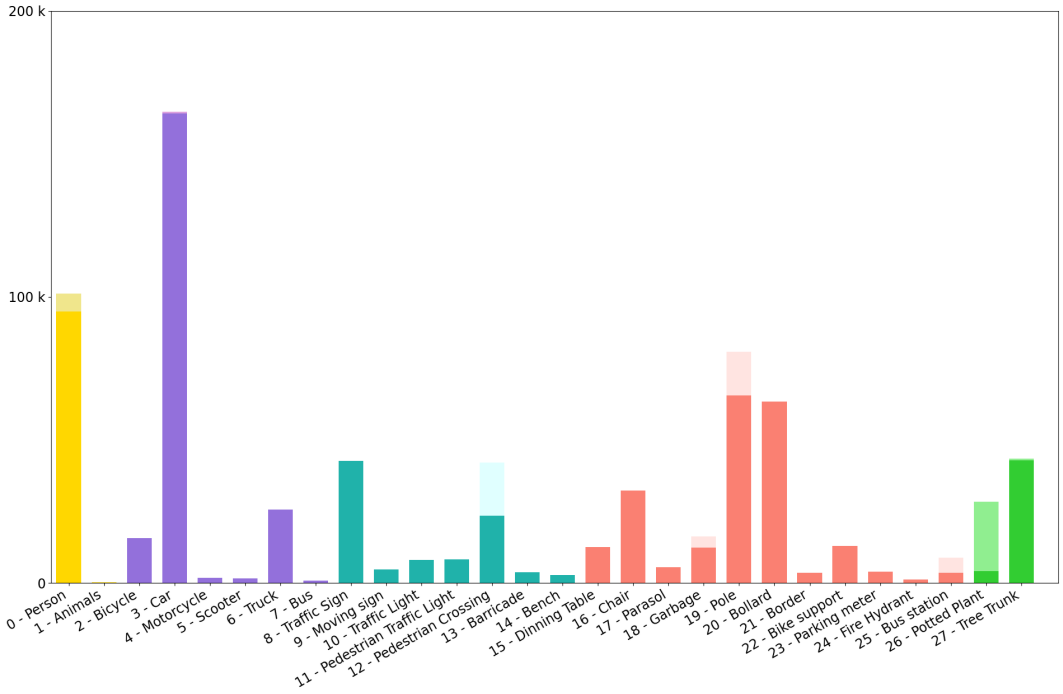


FIGURE 4.7 – Distribution du nombre d'annotations par classe sur l'ensemble du jeu de donnée. Les couleurs pastel sont utilisées pour indiquer la distribution des étiquettes sur les images synthétiques, tandis que des teintes plus prononcées sont employées pour représenter les annotations faites sur les images réelles.

Cependant, le nombre d'apparitions d'un élément différent d'une classe à l'autre en fonction du cadre d'acquisition. Une terrasse de café contiendra principalement des tables et des chaises, tandis qu'un parking aura un nombre significatif de voitures par rapport à d'autres éléments. Ces disparités dans la fréquence des occurrences des différentes classes d'objet sont illustrées dans le graphique de la figure 4.7. La couleur de ce graphique indique un groupe d'éléments partageant des caractéristiques similaires, et les nuances de couleurs distinguent les annotations sur les données synthétiques (couleurs pastel en haut des colonnes) des données réelles (couleurs vives en bas). Selon le graphique, les personnes et les véhicules forment une large part des annotations, alors que d'autres, comme les bornes incendie, sont moins fréquemment représentés. Néanmoins, les catégories les plus dangereuses ou essentielles pour les personnes, notamment comme les objets dynamiques et les panneaux de signalisation, prédominent dans cet ensemble de données. Cette particularité confère une valeur et un intérêt significatifs à ce jeu de données pour l'apprentissage de méthode de détection d'objet.

4.1.2 Détection des obstacles par apprentissage supervisé

4.1.2.1 Base d'apprentissage

Le processus d'apprentissages des méthodes de détection d'objets ou par extension d'obstacle dans certaines situations nécessite des données étiquetées en lien avec les obstacles couramment rencontrés lors d'un déplacement piéton dans un espace urbain. La généralisation et la robustesse de ces méthodes dans différents contextes environnementaux reposent principalement sur l'architecture choisie et la diversité des données d'apprentissage. Une base d'apprentissage riche, comprenant une variété de conditions météorologiques, de niveaux de contraste et de types d'objets, contribue à une détection plus efficace et fiable des obstacles. La compilation de jeux de données provenant de divers environnements augmente non seulement la quantité d'exemples disponibles, mais élargit aussi l'éventail des scénarios que peut rencontrer une personne en milieu urbain. Par conséquent, nous avons intégré dans notre base d'apprentissage les jeux de données uB-Geoloc [162] et SideGuide [133] poussée par leurs complémentarités au niveau des scénarios et des objets annotées.

L'assemblage de ces deux ensembles a donné naissance à un vaste jeu de données de 181 346 images annotées partitionné en trois sous-groupes (entraînement, validation et test) suivant une répartition de 80 %, 10 % et 10 % respectivement. L'harmonisation des annotations entre ces deux ensembles a donné lieu à un regroupement d'éléments similaires pour former 22 classes distinctes. Les objets similaires comme les bornes à incendie et les bornes anti-stationnement ont été regroupés en une seule classe pour augmenter le nombre d'occurrences tout en conservant un intérêt identique pour discriminer le danger engendré par ces objets. Leur identification doit être précise et robuste, mais également rapide afin de permettre à une personne malvoyante assistée par une méthode de substitution sensorielle de localiser et d'interpréter un obstacle environnant. Cependant, bien que le jeu de données soit un élément essentiel pour les performances d'une méthode d'analyse d'image par apprentissage, celui-ci doit être couplé par un modèle algorithmique et un mécanisme d'apprentissage adapté à la situation.

4.1.2.2 Architecture de détection d'objet et résultat

Les architectures de réseaux de neurones n'ont pas même les caractéristiques en termes de faculté de généralisation d'un problème ou de précision, mais également d'occupation de la mémoire et du temps d'exécution. Des modèles spécialisés de détection d'objet ont été conçus pour favoriser un temps de prédiction temps-réel des données sans

compromettre la précision ou la robustesse. En effet, bien que la précision puisse être en deçà d'architectures plus complexes, leurs performances temps-réel donne accès à une détection des éléments entourant une personne en continu pour améliorer le confort et la sécurité des personnes malvoyantes. En effet, une latence élevée dans l'analyse des images peut être problématique pour l'utilisateur d'un système d'assistance et engendrer des retards significatifs dans l'encodage sonore. Ces retards peuvent créer un décalage entre le champ de vision de la caméra et le signal sonore perçu, rendant la navigation difficile et potentiellement dangereuse. Cependant, une précision trop faible dans la détection des objets peut avoir des conséquences graves sur la sécurité de la personne malvoyante. Si un obstacle n'est pas détecté ou s'il y a confusion entre des éléments de nature différente, cela peut conduire à des situations dangereuses. Par exemple, un objet présentant un danger important peut être interprété à tort comme un élément statique de moindre intérêt, mettant ainsi en péril la sécurité de l'utilisateur.

Architecture	mAP50	Précision	Rappel	Temps (ms)	Mémoire (Mo)
YoloV5s ∼ 640 × 640	79,9 %	85,4 %	71,3 %	6,29	474
YoloV5m ∼ 640 × 640	85,1 %	**87,5 %**	77,3 %	7.72	590
Yolov6m ∼ 640 × 640	83,50 %	86,10 %	75,70 %	12,66	607
Yolov7m ∼ 640 × 640	72,8 %	80,6 %	64,4 %	7,70	548
Yolov8n ∼ 1280 × 1280	75,1 %	77,1 %	68,0 %	**5,61**	**389**
Yolov8m ∼ 640 × 640	**85,9 %**	87,0 %	**78,0 %**	8,80	469

TABLE 4.3 – Comparaison des performances entre différentes architectures de détection d'objet évaluées sur un ordinateur portable (Carte graphique RTX 4050 laptop).

Une brève évaluation des capacités de détection des éléments dans un environnement urbain a été conduit sur plusieurs architectures temps-réel pour comparer les résultats après une phase d'apprentissage sur la base de donnée fusionnée mentionnée précédemment. Les paramètres clés, comme le taux d'apprentissage, ont été finement ajustés pour chaque architecture, afin d'optimiser la généralisation. L'entraînement a été arrêté lorsque les performances sur le jeu de données de validation ne s'amélioraient plus.

Le tableau 4.3 récapitule les résultats obtenus pour les différentes architectures de modèle de détection d'objet temps-réel évaluées sur un ordinateur portable. Parmi elles, l'architecture YoloV8m se distingue par ses performances supérieures, tout en maintenant un temps de traitement réduit. Bien que ses avantages soient modestes par rapport à YoloV5m, l'efficacité de YoloV8m en termes de temps d'inférence rapide est cruciale pour

envisager son intégration dans un dispositif embarqué, destiné à assister la navigation des personnes malvoyantes, avec un encombrement réduit et une consommation énergétique optimisée. Suite à ces résultats, le réseau YoloV8m a été sélectionné pour la suite de notre approche. En termes de mesure de performance supplémentaire, ce modèle a montré une capacité de distinction claire entre différentes classes, classifiant correctement 168 888 sur 170 249 obstacles, résultant en un taux de classification correct de 99,2 %. Le faible score de confusion du modèle réduit considérablement le risque de mal juger le danger posé par les obstacles lors de la navigation avec une Intersection sur Union (IoU) fixée à 50 % de 85,9 %. Les résultats détaillés par classe sont présentés dans le tableau 4.4. Les classes "Scooter" et "Banc" affichent des performances nettement inférieures aux autres classes. Cette situation est attribuée à une capacité de généralisation plus faible de ces classes, probablement due à un volume de données d'entraînement moins conséquent pour ces catégories spécifiques.

Parallèlement, les erreurs affectant principalement la précision et le rappel sont dues aux difficultés de détection d'objets éloignés et aux reconnaissances incorrectes en arrière-plan. La résolution de l'image d'entrée, bien qu'adéquate pour détecter des objets proches de l'utilisateur, s'avère insuffisante pour reconnaître des éléments très éloignés, dont la taille apparaît réduite en raison de la perspective. L'intégration de données de profondeur en complément des informations couleur contribue à réduire cette erreur et ainsi à améliorer la robustesse de la détection. De plus, les obstacles situés à une distance très éloignée ont un intérêt moindre pour la compréhension d'une personne aveugle de son environnement. Les obstacles, délimités par une boîte englobante et de manière plus spécifique par la position 2D du centroïde sont enrichis par l'intégration d'une information sur leurs éloignements à partir de la carte de profondeur pour former une représentation

Voiture	Camion	Bus	Moto	Scooter	Vélo
96 %	92,3 %	88,8 %	88,8 %	76,4 %	89,6 %

Personne	Tronc d'arbre	Pot de fleur	Poteau lumineux	Poteau	Potelet
87,7 %	86,3 %	82,2 %	84,6 %	87,2 %	83,6 %

Signalisation	Table	Chaise	Panneau mobile	Barrière	Banc
84,6 %	86,9 %	85.7 %	82,6 %	81,9 %	72 %

Poubelle	Parcmètre	Arrêt de bus	Support à vélo
83,2 %	90,6 %	90,4 %	91,5 %

TABLE 4.4 – Précision moyenne avec une intersection sur union (IoU) supérieure à un seuil de 50%.

3D. Cette mise en correspondance du résultat de la détection d'obstacle avec une carte de profondeur est illustrée sur figure 4.8. Sur l'image (A), deux personnes sont détectées, et cette détection est ensuite croisée avec les données de profondeur (B) pour créer une carte de centroïdes modulée selon la distance (C). Cette représentation s'appuie sur le concept de pixel actif, employé pour encoder en signaux sonores les éléments proches d'une personne malvoyante, évoqué dans la section 3.3.

A B C

FIGURE 4.8 – Représentation tridimensionnelle des obstacles détectés à proximité d'une personne malvoyante. **A.** Détection des obstacles environnants via un réseau de neurones dédié à la détection d'objets. **B.** Cartographie de profondeur indiquant la distance des éléments de la scène par rapport à la personne malvoyante. **C.** Modulation des centroïdes des personnes en fonction de la distance.

4.1.2.3 Dangerosité des obstacles

Le niveau de danger de chaque élément dépend essentiellement de ses caractéristiques intrinsèques, telles que sa nature, sa vitesse de déplacement et sa distance par rapport à l'utilisateur. En effet, le déplacement rapide d'un élément dans un espace de navigation nécessite d'avertir une personne aveugle en amont afin de l'éviter ou de le contourner par rapport à un élément immobile. Une hiérarchisation des obstacles, tel que réalisé intuitivement chez la personne voyante, selon le niveau de danger qu'ils représentent, est essentielle pour guider le processus décisionnel d'un système ou d'une personne. En effet, les capacités cognitives d'une personne pour traiter des informations auditives sont limitées à un nombre réduit simultanément. La réduction du nombre d'informations émis en se concentrant sur les éléments les plus primordiaux permet d'éviter une surcharge cognitive entrainant une fatigue d'un utilisateur malvoyant. Les différentes catégories du jeu de données sont classifiées en trois groupes pour refléter ces nuances de dangerosité.

- **Faible :** arbre, végétation urbaine, feu de trafic (piéton et véhicule), poteau lumineux et de trafic, borne anti-stationnement et à incendie, panneau de trafic, table, chair, panneau mobile, barrière, banc, poubelle, parcmètre, station de bus, et support à vélo.
- **Modéré :** personne.
- **Élevé :** voiture, camion, bus, moto, trottinette et vélo.

Les éléments à haut risque dans un environnement urbain sont principalement constitués d'obstacles dynamiques, à l'exception notable des piétons, dont la vitesse de déplacement est généralement moindre. Leurs déplacements représentent une menace pour une personne malvoyante à analyser afin d'avertir la personne si besoin. En complément de la détection d'obstacle, une méthode de suivi temporel est employé pour suivre dans le temps la position des centroides basée sur l'algorithme SORT. L'algorithme SORT ou *Simple, Online and Real-Time tracking* [163] est spécialement conçu pour permettre un suivi temporel en temps réel à partir de deux algorithmes : l'algorithme hongrois [164] et l'algorithme de Kalman [165]. L'algorithme de Kalman est employé pour prédire la position future des objets en mouvement en se basant sur l'historique des positions observées, offrant ainsi une estimation continue et mise à jour de la trajectoire de l'objet. Parallèlement, l'algorithme hongrois, une méthode d'optimisation, est utilisé pour associer les centroïdes détectés par le réseau de neurones d'une image à ceux des images précédentes. Cette association se fonde sur le calcul du coût le plus faible pour apparier les centroïdes entre deux images consécutives, en tenant compte de la distance et des facteurs de similarité permettant de reconstruire la trajectoire d'un objet au fil du temps.

La complexité de l'analyse des trajectoires d'objets en mouvement dans un environnement urbain est accrue par les mouvements propres de la personne malvoyante. Ces mouvements induisent des changements de position et du champ de vue de la caméra et un mouvement relatif de tous les éléments dans la scène qui s'ajoutent aux déplacements des éléments dynamiques. Même si la vitesse de déplacement d'une personne à pied est modérée, les changements brusques d'orientation, particulièrement lorsque le dispositif d'acquisition est situé au niveau de la tête, compliquent davantage la situation. La connaissance de l'orientation de la caméra par l'utilisation de données inertielles offre une solution partielle à cette problématique. En effet, ces données fournissent des indices précieux sur l'orientation relative de l'utilisateur par rapport à son environnement, aidant ainsi à prédire le mouvement des objets environnants. Cependant, ces informations seules restent limitées pour une analyse précise, mais offrent un moyen de discriminer les éléments potentiellement dynamiques ou immobiles de la scène.

4.2 Détermination des zones accessibles

L'estimation de la dangerosité dans un environnement urbain ne se limite pas à la localisation d'obstacles dynamiques ou statiques. Les zones non adaptées aux piétons, telles que les voies de circulation routière, représentent un risque significatif et nécessitent une prise en compte rigoureuse pour assurer une navigation sécurisée. Les dispositifs existants, tels que les chemins podotactiles, sont principalement localisés dans des zones récemment rénovées ou à des points stratégiques comme les arrêts de bus ou les passages piétons, mais manquent souvent de continuité. L'identification de ces zones accessibles aux piétons est essentielle par leurs caractères non perceptibles naturellement hormis par une connaissance antérieure de cet espace de navigation sans information visuo-spatiale. À l'instar des techniques employées pour la détection d'obstacles, l'utilisation de méthodes de vision artificielle s'avère précieuse.

4.2.1 Segmentation sémantique de l'espace

Les méthodes de segmentation sémantique permettent une analyse topographique des données visuelles afin d'extraire et de classifier des zones en fonction de leur pertinence ou de leur dangerosité. En particulier, les méthodes par apprentissages profonds détaillées dans la section 2.3.2.3 permettent de réaliser une cartographie visuelle précise des espaces urbains. En effet, ces méthodes sont capables d'identifier finement les zones susceptibles de présenter un danger pour les personnes aveugles, telles que les routes, comme illustré dans la figure 4.9. Cette figure représente la segmentation sémantique d'un environnement urbain en différentes zones colorées, où chaque couleur représente un type

FIGURE 4.9 – Illustration d'une représentation sémantique d'une image de rue. La couleur rouge symbolise la route, la couleur orange le trottoir et le mauve l'arrière-plan.

d'espace : le trottoir est marqué en orange, la route en rouge, et l'arrière-plan en mauve. Au-delà de cette distinction basique, la segmentation sémantique peut également être utilisée pour identifier des éléments plus spécifiques et pertinents des zones piétonnes, comme les chemins podotactiles ou encore les passages piétons, souvent marqués par un zébrage distinctif.

Cependant, l'intégration d'une méthode de segmentation sémantique dans les systèmes d'assistance pour personnes aveugles doit tenir compte des contraintes similaires à celles rencontrées dans la détection d'obstacles. En effet, la capacité d'estimation rapide du danger d'un espace est également fondamentale. De plus, cette exigence de rapidité est accentuée par l'exécution en simultané de la méthode de détection d'obstacles. Cette cohabitation se traduit par un partage de ressources matérielles computationnelles. Dans ce contexte, un processus exigeant en ressources matérielles influence défavorablement le second processus, impactant ainsi l'efficacité globale du système.

L'emploi d'architectures de segmentation sémantique basées sur l'apprentissage profond, spécifiquement conçues et optimisées pour permettre des inférences temps réelles, permet de répondre à cette contrainte sans sacrifier la précision nécessaire à la segmentation de la scène visuelle. En effet, la robustesse et la précision sont des critères déterminants pour une caractérisation adéquate de la nature du sol, une tâche complexifiée par la diversité des matériaux de revêtement rencontrés dans un environnement urbain. Ces voies peuvent être constituées de divers matériaux tels que des goudrons avec des teintes différentes, du gravier, ou encore en terre. De surcroît, la distinction entre ces différentes surfaces peut parfois reposer sur des marqueurs subtils, comme une simple bande blanche sur le côté de la chaussée.

4.2.1.1 Base d'apprentissage

La myriade de configurations possibles pour définir une voie piétonne au sein d'un environnement urbain requiert l'emploi de larges jeux de données pour couvrir un large spectre de scénarios. La richesse et la diversité des données sont cruciales pour s'assurer que les modèles puissent estimer avec précision des situations inédites non présentes durant la phase d'apprentissage. De plus, la sensibilité des réseaux de segmentation aux biais potentiels des données d'apprentissage liées à la perspective d'acquisition des images est une considération importante. Par exemple, des données principalement acquises dans un contexte urbain, avec un point de vue centré sur le conducteur d'une voiture, comme il est répandu dans le cas du développement de véhicules autonomes, peuvent introduire un biais. Ce biais se manifeste par la présence des zones piétonnes aux extrémités de l'image, tandis que les voies pour véhicules motorisés occupent une

large partie centrale. En effet, l'absence de zone piétonne au centre de l'image est acquise durant l'apprentissage et se répercute sur les prédictions. L'utilisation de données complémentaires avec un point de vue différent est ainsi nécessaire comme les jeux de données SideGuide ou Mapillary. Le premier jeu de donnée, précédemment employé dans le cadre de la détection d'obstacle pour ses annotations d'objet, fournit des annotations d'images axées sur la nature du sol proche en excluant les informations plus éloignées. Mapillary, quant à lui, comprend un vaste éventail d'images annotées, y compris des détails d'arrière-plan, collectées dans divers environnements, urbains ou non, et depuis divers points de vue, piétons ou conducteurs. Leurs combinaisons offrent ainsi un regard centré sur des espaces réservés aux piétons avec des informations plus ou moins lointaines telles qu'une personne les visualiserait lors d'un déplacement pour définir sa trajectoire. Après l'agrégation, l'ensemble des données, que nous avons obtenus, se composait de 91 399 images annotées et segmentées (Sideguide : 66 399 et Mapillary : 25000) en cinq classes distinctes : route, trottoir, chemin podotactile, passage piéton, et l'arrière-plan pour les éléments extérieurs.

4.2.1.2 Architecture de segmentation sémantique par CNN

L'architecture du modèle est un levier important pour améliorer la robustesse tout en limitant le temps de prédiction. Similairement à la méthode de détection d'obstacle, une évaluation des capacités des architectures de segmentation sémantique temps réel a été réalisée pour définir la mieux adaptée pour le contexte d'utilisation. Les résultats obtenus après une phase d'entraînement et d'évaluation, menée sur six architectures différentes, sont présentés dans le tableau 4.5. Ces résultats, dans leur ensemble, indiquent une qualité de segmentation satisfaisante, tant en termes de précision qu'en termes de rapidité, pour des images d'entrée d'une résolution de 1024×512 pixels.

Au sein de l'ensemble des architectures évaluées, DDRNet [170] s'est distinguée par ses performances supérieures, décrites en détail dans le tableau 4.6 et illustré dans la figure 4.10. Ce réseau CNN a été sélectionné pour la suite de cette thèse pour segmenter les zones accessibles dans les espaces urbains. Cette figure présente quatre exemples visuels de prédictions de segmentation sur l'image originale, où la couleur verte indique la route (zones dangereuses) et le rouge l'arrière-plan. Les zones accessibles aux piétons sont marquées en mauve pour les trottoirs, en gris pour les chemins podotactiles, et en jaune pour les passages piétons. Les performances supérieures de ce réseau étaient attendu par le fait qu'il a été initialement élaboré pour la segmentation de voies en temps réel, un élément central dans le développement des véhicules autonomes. La segmentation des voies routières est une tâche spécifique, similaire à une segmentation de voie piétonne,

Architecture	Backbone	Jutesse	mAP	Temps (ms)	Mémoire (Mo)
BiSeNeT [166]		92,25 %	83,00 %	23,13	422
BiSeNeTv2 [167]		89,72 %	75,93 %	11,19	**289**
STDC 2 [168]		88.44 %	74,09 %	13,23	348
PPLiteSeg [169]	STDC 2	89,62 %	76,36 %	15,10	347
Mobileseg [120]	Mobilenet	91,58 %	81,56 %	**9,3**	317
DDRNet [170]		**92,30 %**	**85,61 %**	11,57	392

TABLE 4.5 – Comparaison des performances entre différentes architectures de segmentation sémantique par apprentissage profond évaluées sur un ordinateur portable (Carte graphique RTX 4050 Laptop).

caractérisée par le fait que les zones de pixels de même étiquette occupent souvent de grandes étendues de l'image. Cet aspect est illustré dans la sous-figure D, qui représente une vaste zone piétonne. Cette particularité de grands espaces contigus permet d'employer des décodeurs de moindre complexité avec des dimensions de carte de caractéristiques faibles combinée à des méthodes de redimensionnements. En effet, le niveau de détail, particulièrement au niveau des contours, a un impact relativement limité sur les métriques de segmentations, contrairement à d'autres applications de segmentation où la distinction précise entre objets adjacents est primordiale. Ainsi, le réseau convolutionnel DDRNET exploite un encodeur performant associé et un décodeur réduit complexité. Les limitations de cette architecture, notamment dans la précision des contours, sont mises en évidence dans les images A et B de la figure 4.10, où la distinction entre la route et le trottoir, ainsi qu'entre la route et le passage piéton, est moins marquée.

Dice	Arrière plan	Route	Trottoir	Pas. piéton	Chemin podo.
92,23 %	86,17 %	87,37 %	81,22 %	86,24 %	87,06 %

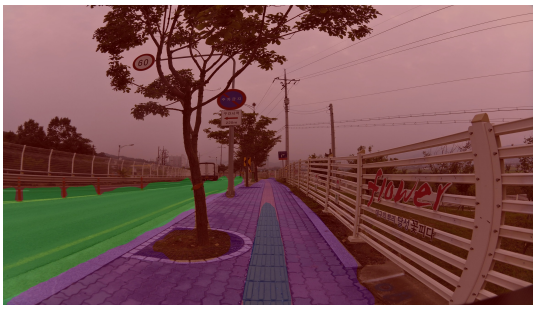
TABLE 4.6 – Précision du modèle de segmentation des piétons évaluée sur une fusion des ensembles de données Mapillary et SideGuide.

Bien que les prédictions soient bonnes pour distinguer de larges espaces, cette faiblesse à la frontière entre deux zones peuvent avoir un impact négatif pour la sécurité d'une personne malvoyante navigant à la limites d'une zone accessible. Dans notre approche, nous appliquons un traitement à base d'opérations morphologiques pour restreindre la superficie des zones accessibles en réduisant leurs contours. Cette technique est complétée par un seuillage appliqué aux pixels ayant une probabilité semblable d'appartenance à

deux classes distinctes. Le seuillage favorise ainsi une sécurité renfoncée pour la personne malvoyante des zones sur lesquelles la distinction entre les classes est moins claire. Une deuxième faiblesse moins impactante se manifeste lors de la segmentation de fines zones lointaines comme le chemin podotactile de la figure 4.10. En effet, sur cette image, la segmentation du chemin podotactile est confondue avec le trottoir à partir d'une certaine distance. Cependant, la segmentation des zones lointaines à un intérêt limité pour une navigation piétonne où la vitesse de déplacement reste modérée.

A. Trottoir pavé

B. Intersection



C. Chemin podotactile

D. Rue piétonne

FIGURE 4.10 – Exemples de prédiction de voie accessible pour une personne malvoyante. La couleur verte (route) et rouge (arrière-plan) représentent des voies dangereuses. Les voies sûres sont désignées par la couleur mauve (trottoir), jaune (passage piéton), et gris (chemin podotactile)

4.3 Conclusion

Un déplacement sécurisé dans un environnement non familier comme un espace urbain requiert une connaissance élevée de notre entourage, en particulier des éléments ou zones dangereux, afin d'en tenir compte pour les éviter. Une perte d'acuité visuelle élevée ou totale empêche de percevoir ces informations cruciales et augmente significativement le sentiment d'insécurité. La compensation d'information non perceptible naturellement par des mécanismes de vision artificielle par ordinateur permettent de compenser en partie

ce manque d'information cruciale. Dans ce chapitre, nous avons exploré des approches basées sur des méthodes de vision artificielle pour améliorer la perception qu'une personne malvoyante peut avoir de son environnement, en se focalisant particulièrement sur l'identification des dangers proches. Nous avons également pris en compte des facteurs clés tels que la robustesse et la rapidité de détection des dangers dans la conception de ces méthodes, afin de les rendre pratiques et efficaces pour une utilisation en situation réelle.

La première approche que nous avons proposée s'est intéressée à l'extraction et à l'analyse des éléments environnants. Un réseau de détection d'objets a été employé pour identifier et localiser des obstacles statiques et dynamiques dans un périmètre étendu pour offrir ainsi aux personnes malvoyantes une meilleure compréhension de leurs espaces. Cependant, cette approche basée sur un mécanisme d'apprentissage profond nécessite l'utilisation de grandes quantités de données annotées pour une généralisation efficace et adaptée à l'application souhaitée. Dans cette optique, nous avons créé un nouvel ensemble de données spécifique aux obstacles urbains et centré sur les espaces denses d'un centre-ville, entremêlant des espaces partagés ou non avec des modes de transport motorisés [162]. Cette base de donnée a été élaborée pour renforcer la détection d'obstacles propres aux environnements urbains et pour contribuer au développement de futures applications d'assistance ou autres dans des contextes piétonniers. Parallèlement à la détection d'obstacles, nous avons intégré une méthode complémentaire visant à identifier les zones inaccessibles aux piétons. Cette méthode basée sur une segmentation sémantique permet de restreindre les déplacements des personnes malvoyantes vers des zones sûres, réduisant ainsi les risques associés à la navigation dans des espaces potentiellement dangereux.

L'analyse de l'espace visuospatial ouvre la voie à l'élaboration d'un système d'assistance à la mobilité dans des espaces sécurisés bien que non familiers. De plus, l'intégration d'une connaissance topologique supplémentaire du milieu de navigation en complément de ces approches de vision artificielle permet de définir de nouveaux systèmes de guidage sonore. Ces systèmes, analogues à ceux décrits dans le chapitre 3 pour la navigation en intérieur, sont capables d'ajuster dynamiquement les instructions de navigation en réaction aux changements dans l'environnement et à la détection d'obstacles à proximité. Ces approches donnent aux personnes malvoyantes la possibilité de naviguer de manière plus autonome dans des environnements urbains complexes, en réduisant leur dépendance à l'assistance humaine et en augmentant leur confiance et leur indépendance dans leurs déplacements quotidiens.

5

Navigation dans un environnement urbain guidé par des signaux sonores

Sommaire

La navigation dans un milieu extérieur est une action stressante pour une personne malvoyante, qui s'avère très difficile lorsqu'il s'agit de rejoindre une destination sans connaissance préalable sur l'environnement de navigation. Les méthodes conventionnelles telles que les systèmes GPS et les cartes, largement répandues parmi les personnes voyantes, se révèlent moins efficaces pour les malvoyants en raison d'un manque d'informations visuelles pour se repérer dans l'espace et d'identifier les éléments dangereux, statique ou dynamique, pouvant entraver son déplacement. Des méthodes d'analyse structurelle de l'environnement par vision artificielle évoquée dans le chapitre 4 permettent de compenser en partie ce déficit de connaissance et de détecter et de localiser la présence d'obstacle. Une intégration de ces informations visuelles avec une connaissance de la position de l'utilisateur et de l'espace de navigation ouvre la voie à des méthodes d'assistance plus avancées et adaptées avec des mécanismes d'interaction avec l'utilisateur malvoyant adapté au contexte applicatif.

Les méthodes de substitution sensorielle visuelle vers auditive tendent à apporter les informations efficacement à travers l'émission de brefs stimulus sonores. En effet, un encodage sonore, adapté aux capacités humaines, offre une compréhension et la localisation rapide et précis d'un ou plusieurs éléments présents dans l'espace environnant de l'utilisateur. Ces approches recréent une représentation spatiale dans un espace sonore tridimensionnel comme détaillé dans le chapitre 3. Néanmoins, la mise au point d'un dispositif d'assistance efficace pour une navigation dans un environnement dynamique doit tenir compte de facteurs tels que l'ergonomie, la robustesse en conditions difficiles, la rapidité d'analyse et de production sonore, ainsi que de l'autonomie nécessaire pour les longs trajets.

Ce chapitre se concentrera sur la présentation d'un système d'assistance à la mobilité en environnement extérieur, développé et illustré à la figure 5.1. Ce dispositif repose sur une méthode de substitution sensorielle auditive fournissant des indications sur l'orientation à suivre pour parvenir à une destination spécifique, ainsi que sur la détection des obstacles potentiels. Ces éléments, déduits à partir d'informations visuelles et spatiales, décrivent une trajectoire adaptée et sécurisée pour un utilisateur malvoyant. La première partie du chapitre se focalise sur la méthodologie de localisation et d'orientation de l'utilisateur à travers un itinéraire sûr à partir de l'intégration d'information topologique et visuelle. La section suivante présente le dispositif expérimental et ses mécanismes d'optimisation. Ces optimisations visent à assurer des performances d'analyse temps réel avec de faibles latences sur une cible matérielle avec consommation énergétique modérée, tout en évitant la dépendance à des ressources externes pour l'analyse et le traitement des données. Enfin, nous analysons les performances du système lors de différentes démonstrations de

navigation réelle.

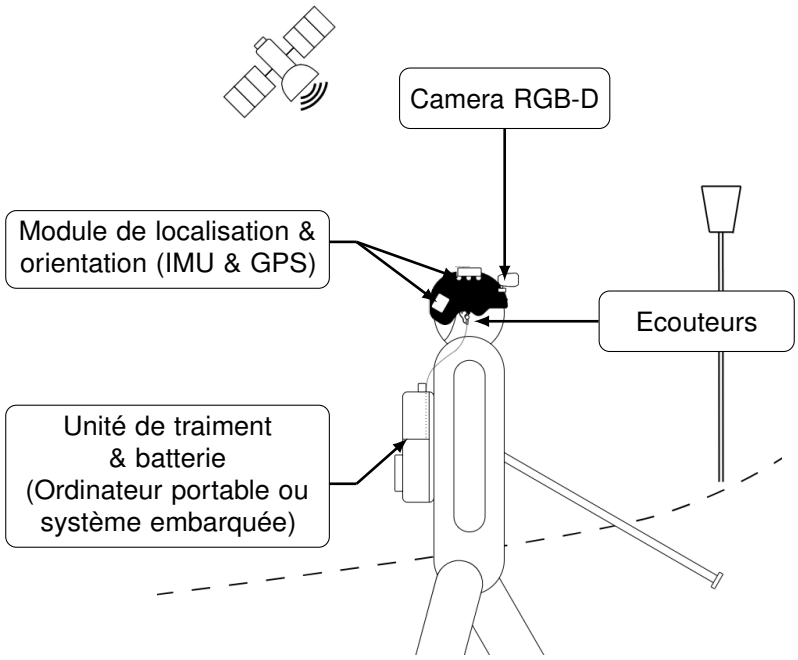


FIGURE 5.1 – Vue schématique du dispositif d'assistance à la navigation.

5.1 Orientation dans un environnement urbain

L'espace urbain représente un milieu complexe et particulièrement hostile pour les personnes aveugles, en raison de la présence d'éléments mobiles dangereux et des difficultés d'adaptation de cet espace pour en améliorer l'accessibilité. En effet, contrairement à un bâtiment où l'ajout d'indications visuelles est réalisable pour définir des informations de localisation et d'orientation adéquates pour naviguer vers une destination, l'espace extérieur requiert une adaptation à sa structure. L'intégration de solutions technologiques telles que l'utilisation de données inertielles et de systèmes de localisation GPS offre un moyen de définir la localisation géographique et l'orientation à suivre si elles sont associées à une connaissance de la topologie de l'environnement et visuel. Ces informations supplémentaires permettent de concevoir des itinéraires optimisés et sécurisés, adaptés pour les personnes malvoyantes, en évitant les zones potentiellement dangereuses.

5.1.1 Détermination d'un itinéaire adapté

La détermination d'un itinéraire adapté pour les personnes aveugles dans un espace urbain exige une approche, tenant compte des multiples variables et des obstacles inhérents à cet environnement. Un facteur réside dans l'analyse et dans l'interprétation de la structure topologique, permettant d'identifier les chemins les plus accessibles et sécurisants. Cette démarche implique non seulement la connaissance de l'agencement urbain des voies et intersections, mais également une connaissance du degré d'accessibilité pour une personne malvoyante des chemins. Des services numériques populaires de cartographie telles que Google Map, Bing Map ou Apple Map permettent un accès à cette connaissance topologique afin de définir la trajectoire. Néanmoins, les fonctionnalités destinées à la navigation piétonne restent limitées sur ces outils et donnent un accès restreint sur des attributs de l'espace de navigation. En effet, les zones urbaines comportent souvent des secteurs dangereux pour les personnes ayant une déficience visuelle, qui doivent être pris en compte pour assurer une navigation en toute sécurité. D'autres services cartographiques, tels que la base de données open-source OpenStreetMap, fournissent une connaissance élargie sur les caractéristiques des voies de circulation piétonne ou non.

La cartographie OpenStreetMap, constamment mise à jour pour s'adapter aux évolutions urbaines telles que les nouveaux développements, les rénovations de bâtiments ou les modifications de voies, fournit des informations essentielles sur les caractéristiques des routes et des sentiers. Cette base de données va au-delà des informations standards telles que les noms de rues, les intersections et les coordonnées géographiques, en incluant des détails vitaux pour la sécurité. La présence de routes réservées aux véhicules motorisés ou de zones spécifiquement piétonnes, telles que les trottoirs, sont associées, facilitant ainsi la détermination de leurs accessibilités pour un déplacement pédestre. L'extraction des informations caractéristiques des différents axes de circulation telles que leurs degrés d'accessibilités et de danger produisent ainsi un filtrage et une pondération pour favoriser un itinéraire sûr et adapté aux piétons, et évitant les zones à risques sans espaces dédiés à la marche à pied. Cependant, des informations pertinentes pour certaines voies peuvent être manquantes et requièrent de définir leurs dangerosités sur d'autres facteurs discriminants comme la vitesse maximale autorisées de circulation comme illustrée à la figure 5.2. En outre, OpenStreetMap, qui s'appuie sur les contributions d'organismes gouvernementaux et de volontaires, peut occasionnellement contenir des erreurs d'étiquetage. Ces erreurs se retrouvent généralement sur des voies moins utilisées, telles que des routes privées ou des sentiers pédestres. Ces inexactitudes constituent une

faible menace pour la sécurité des piétons en raison des caractéristiques inhérentes à ces chemins.

Cette figure représente une cartographie des voies de circulation dans une zone urbaine, où les voies de circulation sont classifiées selon leur niveau de dangerosité. Cette classification repose sur une évaluation à partir de différents critères tels que la vitesse maximale autorisée, le type de route (principale, secondaire, etc.) et la présence de passages pour piétons. Sur cette carte, les routes noires et larges symbolisent les voies interdites ou peu sûres pour les piétons, tandis que les routes représentées par des lignes gris clair et plus fines sont considérées comme étant plus sûres et donc mieux adaptées à la navigation piétonne. Ainsi, les voies noires représentent principalement les voies prohibées pour une navigation piétonne comme les voies rapides ou denses en périphérie des villes. À l'opposé, les voies plus accessibles tendent à se situer au centre-ville ou dans des zones résidentielles ou la vitesse de circulation est moins élevée, le trafic moins dense et les voies réservées aux personnes piétonne plus nombreuse.

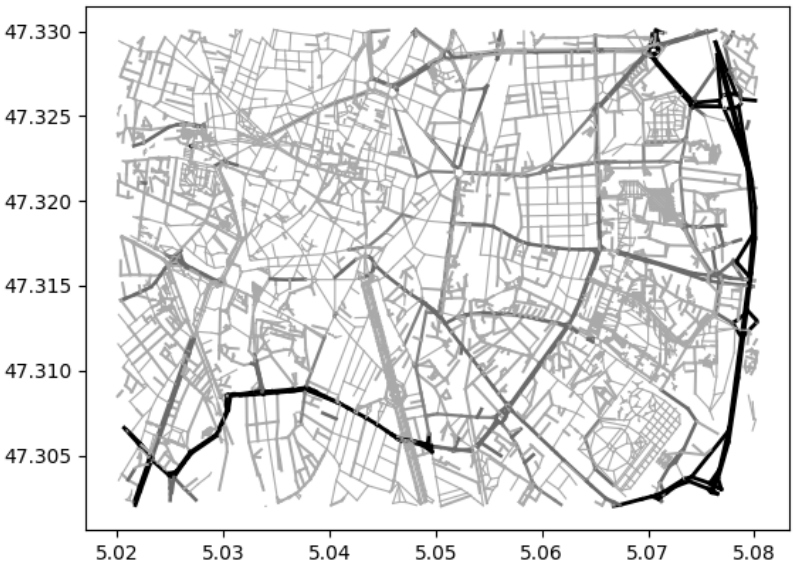


FIGURE 5.2 – Représentation cartographique 2D montrant le niveau de danger des différentes routes dans une ville selon le type de route ou la présence d'un trottoir, les nuances de gris plus claires indiquent les routes à faible danger et les zones noires les plus dangereuses.

En outre, l'exclusion des voies de circulation jugées trop dangereuses, ainsi que des éléments non pertinents présents dans la base de données cartographique, comme les bâtiments, contribue à diminuer considérablement la taille du fichier. En effet, le fichier cartographique complet de la région Bourgogne - Franche-Comté est réduit de 4 Go à 85

Mo en conservant uniquement les voies accessibles aux piétons. Cette réduction permet une diminution du temps de traitement, résultant d'une quantité d'informations à traiter plus restreinte. Ce filtrage cartographique a été réalisé par un outil associé à OpenStreetMap nommée *Osmosis* en conservant uniquement les voies de circulation avec les attributs caractéristiques suivants :

- **Type de route ~ Highway** : Secondaire, Tertiaire, Piétonne, Service (chemin interne), Résidentiel, Escalier, Forestier, Chemin.
- **Piéton ~ Footway** : Trottoir, Passage piéton.
- **Trottoir ~ Sidewalk** : Droit, Gauche, Les deux.

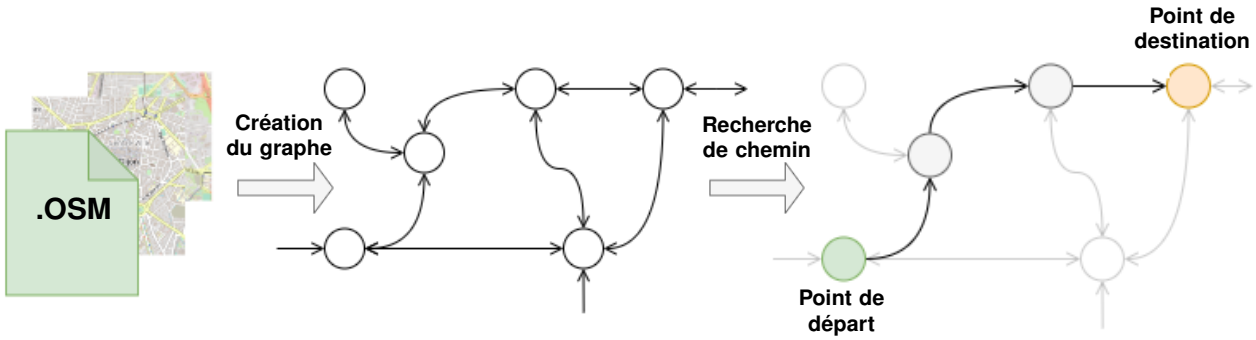


FIGURE 5.3 – Le schéma de la transformation d'une carte OpenStreetMap en une représentation graphique suivant un chemin vers une destination spécifique via des sommets intermédiaires.

Le fichier cartographique optimisé, après avoir filtré pour ne conserver que les voies accessibles, est structuré autour de deux éléments principaux : les *nœuds* et les *voies*. Les nœuds correspondent à des emplacements géographiques clés, comme les intersections ou les courbures d'une route, tandis que les voies désignent les routes elles-mêmes, chacune étant définie par une série de nœuds et leurs informations caractéristiques associées. Similairement à la méthode employée pour la preuve de concept d'assistance à la navigation au sein d'un bâtiment présentée précédemment, cette architecture topologique permet une représentation sous la forme d'un graphe pondéré. Cette structure optimise l'utilisation de méthodes d'estimation de trajets, permettant de tracer un itinéraire à partir d'un point de départ spécifiquement choisi comme illustré à la figure 5.3. Ce point initial dans le graphe est sélectionné pour sa plus grande proximité avec la position initiale de l'utilisateur. Néanmoins, dans des environnements vastes comme un espace urbain, la procédure de sélection du nœud de départ ne peut pas se baser sur une exploration exhaustive de l'ensemble des nœuds du graphe. En effet, la recherche du nœud le plus proche parmi des dizaines de milliers de nœuds dans un espace étendu ou densément

peuplé ralentirait considérablement la méthode. Une réduction de l'espace de recherche à un voisinage de quelques centaines de mètres au lieu d'un unique espace de plusieurs dizaines de kilomètres a été employée pour réduire significativement le nombre de points à parcourir pour définir le nœud de départ, rendant le processus plus efficace et rapide.

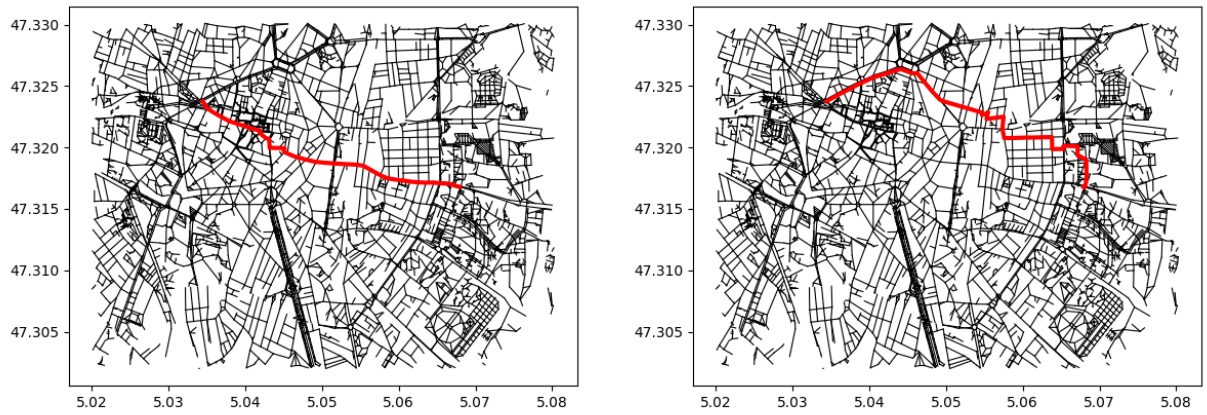
$$w = K \times distance \tag{5.1}$$

La détermination de l'itinéraire à suivre pour rejoindre une destination souhaitée à partir du nœud de départ préalable défini doit être adaptée au profil d'un utilisateur aveugle sans repère visuel pour percevoir la présence de zone dangereuse. Le chemin prédéfini doit être un compromis entre un chemin court et un chemin extrêmement sécurisé afin d'obtenir un chemin personnalisé pour une personne malvoyante. En effet, un court chemin peut être composé d'éléments dangereux et rendre le déplacement difficile, voire impossible. À l'inverse, un parcours plus long, mais plus sûr, peut prolonger le temps de déplacement et solliciter davantage les capacités cognitives de l'utilisateur, entraînant une fatigue inutile. Ces considérations de distance et de dangerosité sont intégrées par une pondération des liaisons dans le graphe, prenant en compte les difficultés topographiques. Ainsi, un long axe de circulation ou trop complexe pour une personne malvoyante se traduira par une valeur accrue. Cette pondération est définie par l'équation 5.1, où K désigne le coefficient de pénalité lié au danger, et *distance* correspond à l'écart entre deux points géographiques, calculé selon la formule de Haversine définie par l'équation 5.2 avec R le rayon de la terre. Cependant, d'autres formules plus simples pour mesurer une distance géographique entre deux points relativement proches peuvent également être adéquates. La figure 5.4 montre la comparaison entre le chemin le plus court et un itinéraire optimisé pour une personne malvoyante, généré par un algorithme de recherche de chemin. Cet itinéraire, bien qu'étant plus long, privilégie des zones réservées aux piétons dans l'espace de navigation, évitant ainsi des voies plus incertaines et potentiellement moins accessibles.

$$\begin{aligned} a &= \sin^2\left(\frac{\Delta lat}{2}\right) + \cos(\text{rad}(lat_A)) \cdot \cos(\text{rad}(lat_B)) \cdot \sin^2\left(\frac{\Delta lon}{2}\right) \\ c &= 2 \cdot \arctan 2\left(\sqrt{a}, \sqrt{1-a}\right) \\ d &= R \cdot c \end{aligned} \tag{5.2}$$

Cette approche diffère de celle utilisée lors de l'élaboration du dispositif d'assistance à la navigation en milieu intérieur, où le graphe était non pondéré et comportait un

nombre limité de nœuds. À l'opposé, la représentation par graphe, symbolisant la structure topologique d'un environnement urbain, est pondérée avec un nombre important de nœuds. Cette architecture nécessite des méthodes de recherche plus sophistiquées telles que l'algorithme de Dijkstra [171] ou ses variantes comme l'algorithme A* [172], pour minimiser le temps de calcul du chemin à suivre comparé à des méthodes de recherche exhaustive. L'algorithme de recherche A*, tenant compte d'une heuristique de distance euclidienne pour le calcul du cout d'un chemin par rapport à Dijkstra dans notre approche pour son exécution plus rapide. La présence d'une heuristique supplémentaire sur la position géographique cible améliore le temps de traitements de l'itinéraire en privilégiant l'étude des voies les plus pertinentes. L'approche implémentée dans notre système d'assistance pour définir l'itinéraire à suivre fut l'algorithme A* afin de fournir une réponse rapide à l'utilisateur aveugle et d'éviter les longs délais d'attente pouvant se traduire par un sentiment de solitude.



a. Chemin le plus court. **b.** Chemin adapté.

FIGURE 5.4 – Comparaison entre le chemin le plus court et le chemin adapté pour une personne malvoyante pour atteindre une destination souhaitée.

L'algorithme de recherche de chemin génère une séquence ordonnée de nœuds, chacun correspondant à l'ensemble des points de passage intermédiaire, avec leurs attributs, nécessaire pour atteindre la destination en débutant par le point de départ. Par conséquent, pour atteindre la destination prévue, la personne malvoyante doit passer par tous les marqueurs géographiques intermédiaires représentant des endroits critiques dans un environnement urbain tels que les intersections ou des courbures de la voie de circulation. Cependant, la simple connaissance de l'itinéraire, même adapté, n'est pas suffisante sans information sur l'orientation et localisation de l'utilisateur pour atteindre la

destination.

5.1.2 Suivi d'une trajectoire

L'absence de repère visuel associé à une connaissance visuelle de son environnement proche empêche une personne aveugle de suivre un itinéraire défini. Ce déficit d'information peut être surmonté par l'acquisition de données complémentaires sur l'environnement de navigation. Le parcours initial est ainsi agrémenté d'informations continues sur la position et l'orientation de l'individu, ainsi que sur l'agencement de l'espace alentour, pour établir une trajectoire sûre et adaptée qui mène l'utilisateur à une destination souhaitée. Ce processus est présenté dans le diagramme de la figure 5.5 où les informations de la carte OpenStreetMap sont associées avec des informations spatiales et visuelles pour définir la trajectoire. Ce diagramme reprend les principes des approches de vision artificielle évoqués dans le chapitre 4 en ajoutant un processus de détermination du chemin et de la direction à suivre nommée δ.

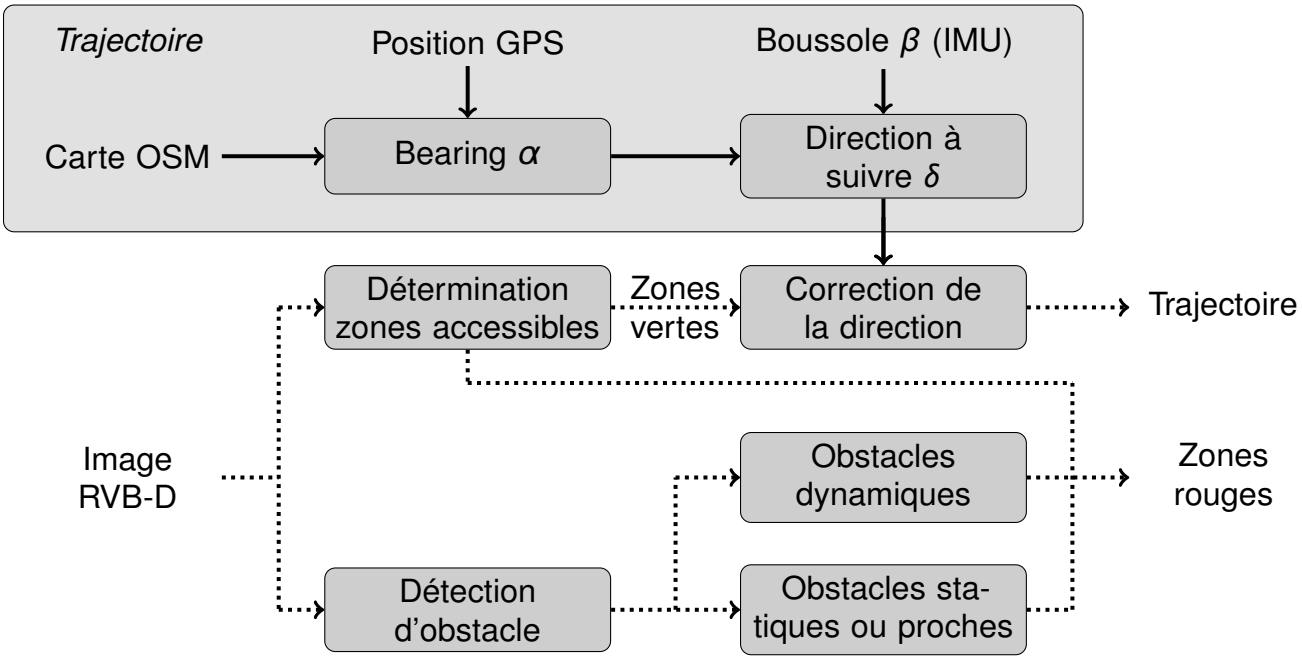


FIGURE 5.5 – Processus de détermination de la trajectoire à partir de donnée spatiale et visuelle.

La direction δ à suivre pour atteindre un lieu désiré, dans le diagramme et la figure 5.6 est déterminée à partir d'une combinaison d'information spatiale acquise par le biais d'un capteur inertiel et d'une antenne GPS pour localiser l'utilisateur. En effet, cette information angulaire définie par l'association de l'angle de relèvement α avec l'orientation de l'utilisa-

teur β. L'angle de relèvement, qui est la direction entre deux points géographiques, est calculé à partir des positions respectives de l'utilisateur et de la destination, notées $user$ et $dest$, ainsi que la valeur de latitude θ et de différence de longitude ΔL.

$$
\begin{aligned}
X &= \cos(\theta_{dest}) \times \sin(\Delta L) \\
Y &= \cos(\theta_{user}) \times \sin(\theta_{dest}) - sin(\theta_{user}) \times \cos(\theta_{dest}) \times \cos(\Delta L) \\
Bearing &= \arctan(X, Y)
\end{aligned}
\tag{5.3}
$$

L'angle de relèvement α, qui représente la direction vers une destination particulière en référence au nord géographique, apporte une indication essentielle, mais reste insuffisante. Cet angle renseigne sur la direction vers laquelle se situe la destination depuis un point de référence fixe (le nord géographique) sans analyser l'orientation actuelle de la personne malvoyante. La connaissance de l'orientation de la personne à partir d'une boussole magnétique compense ce manque. Malgré la légère divergence, nommée déclinaison magnétique, entre les repères géographique et magnétique de la boussole terrestre. La relative constance de cette divergence dans un espace réduit telle qu'une ville permet d'affiner l'orientation magnétique au sein d'un espace géométrique. Par conséquent, la direction à suivre δ est définie par la différence angulaire entre l'angle de relèvement α et l'orientation spatiale de l'utilisateur β.

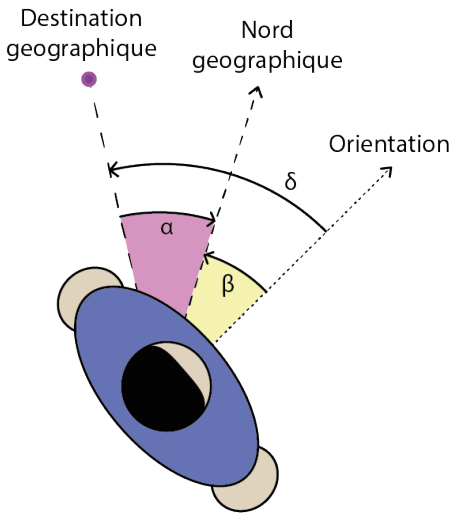


FIGURE 5.6 – Vue schématique de la méthode de détermination de la trajectoire angulaire à suivre pour rejoindre une destination géographique désirée à partir de la position et de l'orientation d'une personne.

Cependant, un itinéraire prédéfini basé uniquement sur les informations spatiales de l'utilisateur et sur la topologie de l'environnement de navigation peut omettre de prendre

en compte les obstacles ou les zones dangereuses pour un piéton, telles que les routes. Le chemin défini pourrait mener à des zones inadaptées ou périlleuses si les directives fournies par le système n'intègre pas des contraintes de l'espace. Par conséquent, l'élaboration d'un itinéraire sécurisé nécessite l'ajout d'informations sémantiques concernant l'espace de navigation. L'espace visuel offre une perception directe de l'agencement de l'environnement et de la présence d'éléments dynamiques susceptibles d'affecter la sécurité lors d'un déplacement. L'exploitation de cet espace au moyen de techniques de vision artificielle permet de définir la présence de zones dangereuses à proximité par des approches de segmentation sémantique. L'analyse sémantique permet d'ajuster l'angle azimutal de l'itinéraire prédéfini vers des zones plus sûres nommées zones vertes, telles que les trottoirs. Cette correction implique de trouver une zone accessible et sécurisée à proximité de la position prévue, que la personne malvoyante puisse atteindre. Dans notre méthode, l'accessibilité est évaluée en recherchant un large espace continu à partir du bord inférieur de l'image, indiquant un chemin accessible à pied sans nécessiter un passage par des zones inadaptées. Le chemin est alors rectifié en sélectionnant, parmi les différentes zones possibles, celle dont le centre (ou centroïde) est le plus proche de l'itinéraire prédéfini. L'analyse géométrique permet de localiser le point central d'une zone ou d'un objet dans l'image, facilitant ainsi l'identification d'un trajet qui suit au mieux la direction souhaitée. L'application de différentes transformations morphologiques de base (érosion, ...) permet, quant à elle, d'éliminer les surfaces non adaptées pour la navigation (non accessible, superficie faible). En privilégiant le centroïde des zones identifiées, on minimise les risques associés à la navigation près des limites potentiellement dangereuses de l'espace piéton, telles que les bords des trottoirs. Cette approche contribue également à réduire les erreurs potentielles liées à la méthode de segmentation des bordures précédemment mentionnée.

5.2 Encodage sonore et apport sémantique

Le guidage sonore relatif à la trajectoire doit être défini avec précision et rapidité par la personne afin d'éviter une mauvaise compréhension l'entraînant vers une zone dangereuse. De plus, l'émission de la direction à suivre peut isoler l'utilisateur de son espace de navigation et réduire son interaction avec son espace sans une perception des éléments environnants tels que les obstacles. Néanmoins, notre preuve de concept d'un dispositif d'assistance à la navigation dans un espace intérieur a démontré une aptitude humaine à percevoir et à interpréter des signaux sonores spatialisés encodant

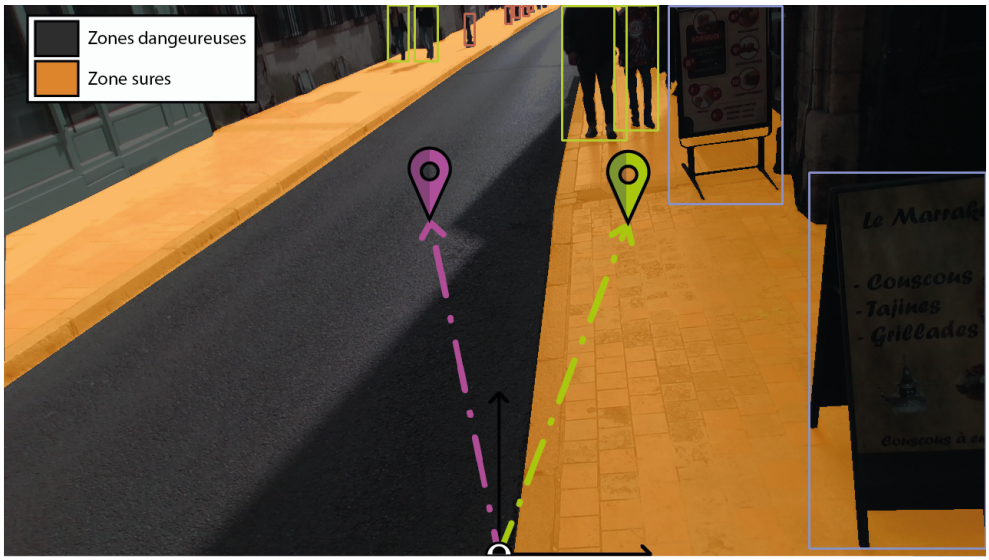


FIGURE 5.7 – Ajustement de la trajectoire à suivre pour une navigation pédestre. La position indiquée en violet est rectifiée vers une voie accessible pour un piéton à la position verte.

ensemble de ces deux éléments cruciaux (Chapitre 3). L'emploi d'une approche similaire et enrichie pour la navigation en extérieur avec la connaissance des obstacles dynamiques nécessitant une détection en amont offre une alternative pour les personnes malvoyantes de naviguer pour rejoindre une destination en toute sécurité dans un espace non familier avec une connaissance de son environnement proche.

Dans le système de navigation dans un bâtiment, les obstacles étaient uniquement déterminés à partir d'un critère de proximité défini par la carte de profondeur sans déterminer leurs natures. L'intégration de cette indication obtenue à partir de l'analyse de l'espace visuelle présentée dans la section 4.1.2 permet de transmettre à l'utilisateur précocement la présence d'un obstacle dynamique et plus tardivement des obstacles statiques comme représenté dans la figure 5.8. Cette figure illustre l'évolution de l'intensité du signal sonore spatialisée en fonction de la nature de l'obstacle. Cependant, la méthode de localisation des obstacles demeure limitée par son nombre restreint d'objets catégorisés (22 classes distinctes) et par d'éventuelles non-détections d'éléments dangereux. L'encodage sonore des données de profondeur est intégré pour pallier ce problème (courbe orange).

La méthode d'association des deux flux sonores relatifs à la trajectoire et aux obstacles environnements définis lors du système de navigation en intérieur est remployé dans le cadre de la navigation en extérieur par sa capacité à transmettre deux informations distinctes en simultanée. Une caractéristique temporelle supplémentaire est ajoutée au signal de la trajectoire pour accroître la capacité de discrimination de ces deux modalités. Le signal sonore indiquant la direction est transmis par intermittence, toutes les 500 millise-

condes, contrairement au signal continu qui sert à détecter les obstacles. En complément, des instructions verbales sont fournies pour guider l'utilisateur lorsqu'il approche d'un nouveau point de repère virtuel (jalon) ou pour lui indiquer la distance restante avant d'atteindre le prochain marqueur, facilitant ainsi sa navigation.

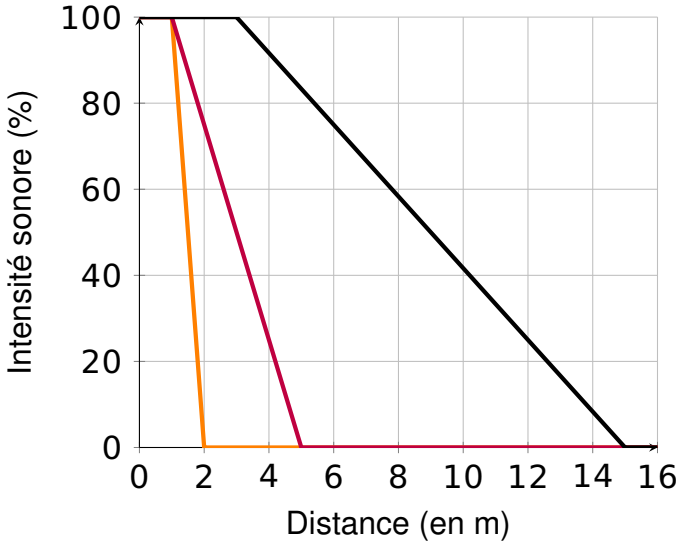


FIGURE 5.8 – Évolution de l'intensité sonore du signal en fonction de la distance d'un obstacle. Les courbes noires et rouges représentent respectivement les obstacles dynamiques et statiques détectés. La courbe orange indique un seuil d'alerte à partir des informations de la carte de profondeur sans connaissance de la nature de l'objet.

5.3 Système

L'intégration au sein d'un dispositif d'assistance à la navigation de cette approche d'orientation dans un milieu urbain ouvre la voie à la réalisation de déplacements libres et autonomes dans des espaces non familiers pour les personnes aveugles, mais doit faire face à des contraintes liées au contexte d'utilisation telles que son ergonomie, son autonomie et sa robustesse. Un dispositif mal adapté pourrait s'avérer être inutile, voire être dangereux pour l'utilisateur. Cette considération est d'autant plus cruciale dans le cas d'une méthode basée sur la substitution sonore, où la clarté et la concision des informations sont essentielles pour garantir une compréhension et une réactivité optimales, assurant ainsi la sécurité de l'utilisateur dans la navigation vers sa destination. Des retards dans la transmission ou des erreurs de perception peuvent entraîner des inexactitudes dans l'identification des obstacles ou la détermination de l'itinéraire approprié. Cette problématique est exacerbée par les étapes d'analyse et de traitement des données

recueillies par les capteurs ou les caméras, ainsi que par l'encodage des informations pertinentes, qui peuvent engendrer des retards significatifs.

En réponse à ces contraintes et considérations critiques inhérentes aux systèmes d'assistance à la navigation pour les personnes malvoyantes, nous avons développé un système optimisé. Ce système, comprenant l'acquisition, l'analyse et la génération sonore, est entièrement autonome et ne dépend pas d'une connexion continue à une source/destination externe, telle qu'une architecture de cloud computing, pour l'extraction de caractéristiques pertinentes. Une telle architecture, dépendante depuis une liaison externe, pourrait compromettre la robustesse du système en cas de perte de signal ou de latence élevée dans la transmission. La solution proposée repose sur une optimisation matérielle et logicielle des divers traitements, assurant ainsi une performance temps-réel.

5.3.1 Description du système expérimental

Le dispositif développé est composé de trois éléments distincts, un dispositif d'acquisition, d'une unité de traitement des informations et d'un dispositif d'émission sonore. L'unité d'acquisition regroupe l'ensemble des capteurs permettant l'obtention des informations spatiales et visuelles essentielles pour la mise en application des approches de détection d'obstacles et de définition de la trajectoire permettant d'atteindre une destination souhaitée. Cet ensemble conçu pour suivre les mouvements de la tête pendant la navigation est situé précisément sur un casque représenté à la figure 5.9. Il se compose des trois principaux éléments d'acquisition suivants, chacun associé à sa propre fréquence :

- Caméra RGB-D ∼ *Intel Realsense D435 RGB-D* ∼ 30 images par secondes
- Antenne GPS (Global Positioning System) ∼ *Adafruit BNO055 IMU* ∼ 10 Hz
- Capteur IMU (Inertial Measurement Unit) ∼ *Adafruit Ultimate GPS breakout* ∼ 100 Hz

La caméra RGB-D, positionnée à l'avant du dispositif d'acquisition, se comporte comme des yeux de substitution pour l'utilisateur. Les informations captées par la caméra fournissent des informations visuelles de couleur et de profondeur sur l'environnement situé devant l'utilisateur. Ces données permettent d'analyser et de déterminer la présence des obstacles ou de zones de navigation incertaines, voire dangereuse pour l'utilisateur. La localisation et l'orientation de l'utilisateur sont déterminées respectivement par une trilatération de sa position par GPS et par des informations obtenus par un capteur inertiel. Cependant, bien que ces éléments répondent aux besoins de notre méthode, des

limites dues à leurs caractéristiques intrinsèques sont à considérer pour appréhender les contraintes de fonctionnement du système. La précision métrique du GPS est significativement affectée par la visibilité des satellites. Dans des zones confinées, comme les centres historiques, cette visibilité réduite peut impacter la précision du GPS, bien qu'elle reste amplement suffisante pour localiser une personne dans des conditions normales, comme dans des espaces ouverts. De plus, l'association avec d'autres méthodes de géolocalisation en complément du GPS tel qu'une triangulation avec les signaux de téléphonie mobile (GSM, LTE, etc) comme utiliser dans nos téléphones mobiles pourrait répondre à cette contrainte et améliorer la précision dans ces espaces. Le dispositif que nous avons développé s'est concentré sur des environnements urbains plus contraints à des espaces ouverts dans le cadre de cette preuve de concept en milieu extérieur.

FIGURE 5.9 – Photographie et description du système d'acquisition.

D'autre part, le capteur inertiel offre une connaissance limitée du déplacement de l'utilisateur, largement provoquée à la présence élevée de bruit sur les données d'accélération. Ce bruit limite l'utilisation des données sur de longues périodes pour préciser la trajectoire suivie ou la vitesse de déplacement. Néanmoins, le capteur inertiel permet de déterminer avec précision l'orientation de la tête de l'utilisateur non seulement par rapport à son point de départ, mais aussi en fonction du référentiel magnétique terrestre. En effet, grâce au magnétomètre inclus, le signal magnétique terrestre peut-être capté afin de déterminer la direction du nord magnétique, mais peut-être perturbé par la présence extrêmement proche d'un élément émettant un champ magnétique élevée ou par l'orientation du capteur par rapport au sol terrestre affectant le vecteur magnétique. Cependant, la disposition des composants sur le casque, éloignée les uns des autres, minimise les perturbations magnétiques. L'impact de la position du capteur sur le vecteur du signal magnétique peut être compensé en utilisant les données d'orientation relative fournies par

le capteur lui-même. L'équation 5.4 calcule la direction du nord magnétique, compensant l'effet de la position du capteur à partir de la rotation du capteur θ et ϕ et du vecteur magnétique M.

$$\begin{aligned} X &= M_z \times \sin(\phi) - M_y \times \cos(\phi) \\ Y &= M_x \times \cos(\theta) + M_y \times \sin(\theta) \times \sin(\phi) + M_z \times \sin(\theta) \times \cos(\phi) \\ Heading &= \arctan(X, Y) \end{aligned} \tag{5.4}$$

Les données acquises sont transmises à une unité de traitement située dans un sac à dos de l'utilisateur à l'aide de deux câbles USB. De plus, ces câbles offrent l'alimentation électrique nécessaire pour le fonctionnement de ces éléments. L'unité de traitement est le cœur névralgique d'une méthode de substitution sensorielle pour générer les signaux sonores relatifs au contexte d'utilisation. Les ressources matérielles doivent être dimensionnées pour permettre un fonctionnement temps réel. Dans cette optique, l'usage initial d'un ordinateur portable offre à la fois une portabilité du système et performances élevées. Le modèle choisi était équipé d'un processeur Intel Core I7 12700h, doté de 14 cœurs et 20 threads, 16 Go de RAM, ainsi que d'une carte graphique RTX 4050. Il possédait également une batterie de 90W/h et fonctionnait sous le système d'exploitation Ubuntu 20.04. Cette configuration permet d'accéder à des méthodes de parallélisation des processus, optimisant ainsi les méthodes d'analyse et de génération sonore pour réduire les délais. Par ailleurs, la disponibilité de bibliothèques en C++ telles qu'Eigen et OpenCV facilitaient l'utilisation de fonctions mathématiques et de traitement d'image optimisées, contribuant à un développement plus rapide et plus efficient.

5.3.2 Architecture multi-processus et GPU

L'optimisation de l'implémentation logicielle de l'approche d'assistance à la mobilité dans un milieu extérieur pour les personnes aveugles a pour objectif principal de réduire le temps nécessaire à la génération de sons pour guider un utilisateur ou l'avertir de la présence d'un obstacle. De plus, l'inclusion d'un vaste espace de navigation avec un nombre substantiel de sommets et d'arêtes est indispensable pour élaborer une méthode de navigation pédestre robuste capable d'exécuter efficacement des algorithmes de navigation dans des environnements denses ou vastes. Une approche réactive qui maximise l'utilisation des ressources disponibles, en mettant l'accent sur l'exploitation d'un système multicœur et d'une unité de traitement graphique (GPU - Graphic Processing Unit), est

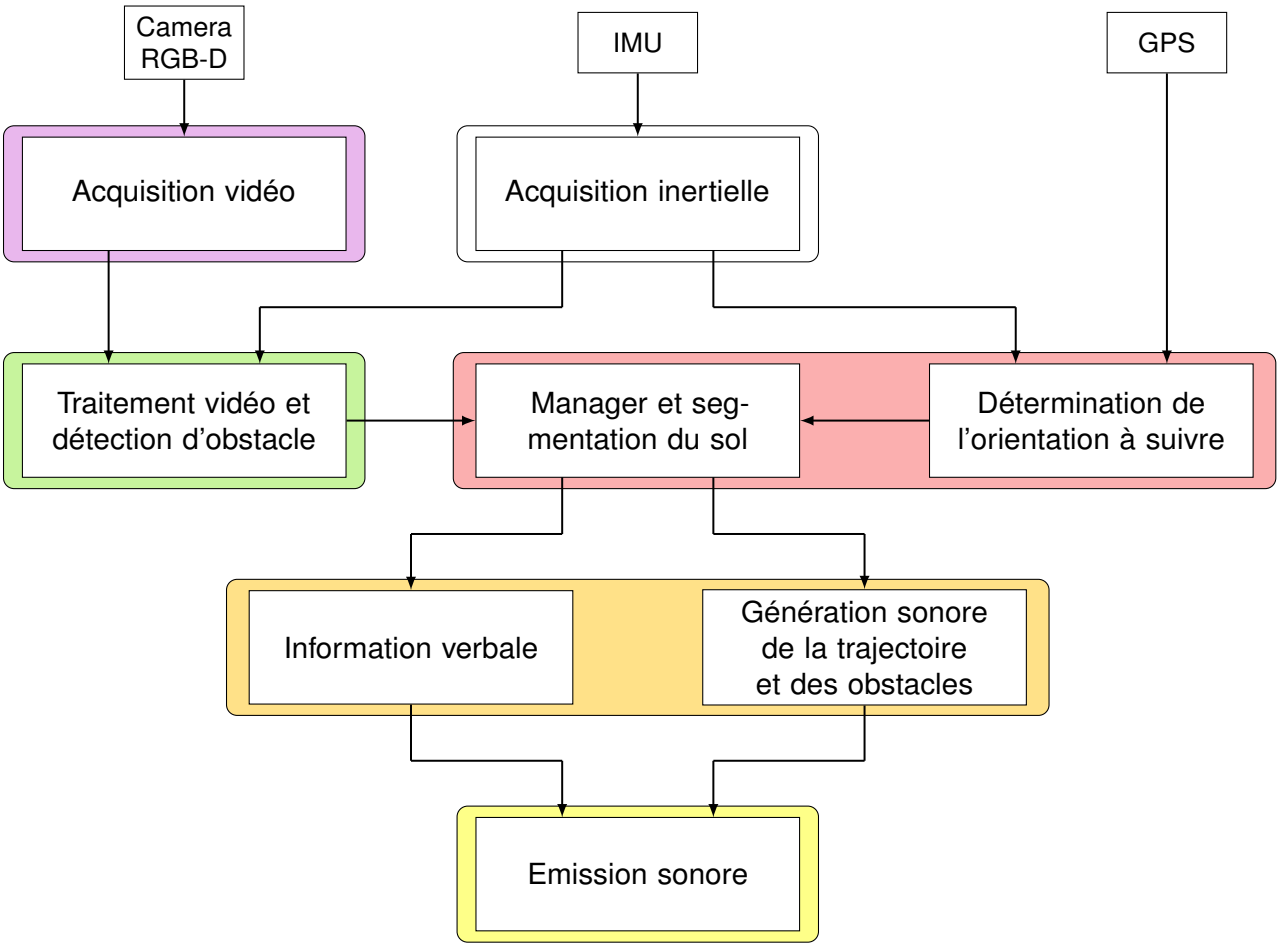


FIGURE 5.10 – Diagramme détaillé de l'architecture du système, depuis l'acquisition des données jusqu'à l'émission sonore. Chaque ensemble de couleurs, regroupant une ou plusieurs fonctions, symbolise un processus (ou thread) dédié.

efficace pour réduire de manière considérable les temps de traitement et d'analyse. Cette méthode permet également une utilisation de ressources matérielles plus limitée, comparée à une gestion moins optimisée du processus de traitement. Dans cette configuration, les processus compatibles sont exécutés en parallèle pour diminuer de manière significative les temps de traitement. En exploitant les capacités de parallélisation de notre méthode, tant pour la détection d'obstacles, l'établissement de la trajectoire à suivre et la génération sonore, nous avons élaboré une approche fondée sur une architecture à la fois multi-processus et CPU/GPU.

Dans cette approche, les différents traitements qui structurent notre méthode de substitution sensorielle de visuel à auditif ont été séparés en processus fonctionnant simultanément pour acquérir ou traiter les données suivantes sans attendre la fin de la génération ou de l'émission du flux sonore des informations précédentes. Cette structure

multiprocessus est illustrée dans la figure 5.10, où chaque processus est illustré par une couleur différente. Ces processus regroupent une ou plusieurs sous-tâches de notre méthode d'assistance à la mobilité et sont orchestrés par un processus central désigné sous le nom de *Manager*. Dans cette architecture, cinq processus secondaires sont affectés à des fonctions spécifiques et distinctes. Plus précisément, deux de ces processus se concentrent sur l'acquisition de données fournies par la caméra et le capteur inertiel, tandis que deux autres sont dédiés à la génération et à l'émission de sons. L'ensemble de ces processus s'exécutant synchronisées aux fréquences d'acquisition ou d'émission sonore. Les traitements d'analyse vidéos pour la détection des obstacles environnants et des zones accessibles sont partagés entre un processus dédié et le processus manager s'occupant de l'estimation de la trajectoire à suivre.

Calcul et correction	**Génération sonore**	
de la trajectoire	**Trajectoire**	**Obstacle**
2 ms	<0.1 ms	0.4 ms

TABLE 5.1 – Temps de traitement de la génération sonore et du calcul de la trajectoire.

L'ensemble de ces processus impacte le délai entre l'acquisition des informations visuelles et spatiales jusqu'à la génération sonores de la trajectoire et des éléments environnants. Le détail des temps de calcul des différents processus sont reportées dans les tableaux 5.1 et 5.2 en contournant les processus d'acquisitions de la caméra et des capteurs pour éliminer l'impact des fréquences d'acquisitions sur la latence. Les résultats obtenus montrent une influence faible des processus de génération sonore ou de détermination de la trajectoire contrairement aux méthodes de traitements vidéos. En effet, les processus de détection d'obstacles et de segmentation sémantique impactent lourdement le temps global, cependant un traitement temps-réel est atteint grâce à l'optimisation multiprocessus proposée et à l'externalisation des opérations de vision artificielle sur le GPU. En outre, l'efficacité des architectures de réseaux de neurones a été améliorée grâce à l'utilisation de la bibliothèque TensorRT de Nvidia, une technologie conçue pour optimiser, accélérer et déployer les modèles d'apprentissage automatique dans un environnement de production. En termes de performance, le processus a atteint une utilisation maximale de 13,70% sur l'ensemble du CPU, avec 4,7 Go de RAM utilisée. L'utilisation du GPU a culminé à 76%, avec une utilisation de 1,337 Go de VRAM et une consommation d'énergie de 74 watts.

	Traitement vidéo		Traitement
	Segmentation	Détection	global
CPU - Libtorch (ms)	1130	142	1136
GPU - TensorRT (ms)	4.9	10.4	14.2

TABLE 5.2 – Comparaison du temps de traitement sur CPU et GPU des méthodes de segmentation et de détection et leurs impacts sur le processus global.

En revanche, la configuration reposant exclusivement sur le CPU s'est révélée insuffisante pour assurer un temps de réponse adapté à une utilisation du système dans un contexte réel. Les délais de traitement, s'étirant au-delà d'une seconde, compromettent l'assurance d'une navigation aisée, fluide et sécuritaire. Ces observations mettent en lumière l'impératif de l'accélération par GPU dans l'application des techniques de vision artificielle, notamment pour la segmentation sémantique, afin de parvenir à une performance optimale du système.

5.3.3 Implantation sur une cible embarquée

L'efficacité de notre approche d'assistance à la mobilité sur un ordinateur portable, équipée d'un GPU, a ouvert la voie au développement d'un dispositif sur une cible matérielle embarquée avec consommation énergique plus modérée pour prolonger l'autonomie. Ce changement de cible matérielle permet une amélioration de l'ergonomie pour un utilisateur mobile, en réduisant le poids et l'encombrement. De tels ajustements favorisent des périodes d'utilisation prolongées, réduisant ainsi le risque d'interruptions inopportunes qui pourraient semer un sentiment d'insécurité chez l'usager. Cependant, il demeure crucial de sélectionner une cible embarquée qui, tout en étant sobre en énergie, supporte une parallélisation des traitements des réseaux de neurones convolutifs.

Diverses architectures matérielles intègrent des modules de parallélisation comme les cibles matérielles ARM/GPU (Advanced RISC Machines) et FPGA (field-programmable gate array). La plateforme ARM/GPU associe sur un même *system on chip* des unités de traitements graphiques et un processeur ARM plus économe en énergie que les processeurs x86 de nombreux ordinateurs personnels. En ce qui concerne les cibles de type FPGA, elles offrent une flexibilité et une personnalisation élevées, bien qu'ils puissent nécessiter généralement des délais d'implantation plus longs malgré l'expérience du laboratoire sur les cibles FPGA [173], [174] en utilisant des approches de prototypage rapide [175], [176]. Nous avons opté pour une cible GPU pour la phase de définition de la

preuve de concept. Cependant, la relative plus faible consommation des cibles matérielles de type FPGA couplé au développement des outils commerciaux ou académiques de prototypages en particulier pour l'implantation de réseaux de neurones [177], [178], permettent d'envisager l'utilisation de ce type de cibles matérielles dans des développements futurs.

La cible matérielle utilisée pour se substituer à l'ordinateur portable est une carte Nvidia Orin AGX 32 Go basée sur une architecture ARM/GPU. Cette plateforme se distingue par une efficacité énergétique supérieure par rapport à l'implémentation sur l'ordinateur portable, tout en conservant une grande partie de l'architecture logicielle préexistante. De plus, elle supporte des outils d'optimisation de pointe comme la bibliothèque TensorRT, essentielle pour l'inférence de réseaux de neurones, assurant ainsi une performance accrue. La plateforme offre également une variété d'options configurables, y compris les modes d'énergie et les paramètres de précision en virgule flottante, influençant directement les temps d'inférence comme indiqué dans le tableau 5.3.

	FP32 MaxN	FP32 50W	FP32 30W	FP16 MaxN	FP16 50W	FP16 30W
Détection (ms)	17	23	46	8	12	26
Segmentation (ms)	9	14	36	4	6	18
Temps global (ms)	27	38	81	20	27	44

TABLE 5.3 – Comparaison du temps de traitement pour la vision et les processus globaux avec différents modes de consommation d'énergie et de dynamique en virgule flottante.

La transition vers une précision réduite, de 32 bits à 16 bits par exemple, diminue la demande en mémoire et augmente la vitesse des calculs grâce à une réduction du volume de données traitées, tout en préservant la précision des modèles ou basées sur des approches d'apprentissage profond. En effet, les modèles ont été entraînés grâce à une technique d'apprentissage utilisant une précision mixte, combinant des valeurs à virgule flottante avec une précision simple (FP32 ∼ *Simple floating point precision*) ou à demi précision (FP16 ∼ *Half floating point precision*). Cette méthode permet de confier les tâches les plus lourdes au FP16, allégeant ainsi la charge de calcul, tout en réservant le FP32 pour les opérations nécessitant une plus grande précision en FP32. Par ailleurs, les Tensor Cores des GPU NVIDIA optimisés pour les calculs en FP16 et INT8 améliorent grandement les performances de calculs par rapport aux opérations en FP32. Quant aux modes d'alimentations, ils correspondent à des réglages prédéfinis par le fabricant de la cible matérielle et spécifient la fréquence et le nombre de cœurs activés simultanément pour limiter la consommation d'énergie. Le système résultant permet un traitement en temps réel avec une faible latence, autorisant ainsi la proposition

d'expériences en extérieur.

Les résultats exposés dans le tableau 5.3 illustrent la capacité du système à atteindre un traitement en temps réel limité par la fréquence d'acquisition de la caméra à 30 images par seconde dans les configurations les plus optimales (mode de puissance maximale et précision de 16 bits). Néanmoins, des configurations moins performantes pourraient s'avérer suffisantes pour qu'une personne malvoyante puisse accéder aux informations nécessaires lors d'une navigation urbaine par substitution auditive. Cependant, la performance de ce système ainsi que des méthodes de vision artificielle requièrent une mise à l'épreuve dans des conditions d'utilisation pour évaluer l'adéquation de l'outil pour une personne malvoyante et pour identifier diverses limitations inhérentes à son utilisation.

5.4 Expérimentation

5.4.1 Protocole expérimental

Nous avons conduit une évaluation de l'utilisation de ce système expérimental d'assistance à la navigation pour les personnes malvoyantes. Cette expérimentation, similaire à celle réalisée précédemment au sein d'un bâtiment, consistait à suivre une personne équipée du système et les yeux bandés dont les informations sonores devait la conduire jusqu'à une destination désirée en toute sécurité. Le participant avait une connaissance préalable de l'ensemble des caractéristiques des signaux sonores spatialisés. Cependant, et afin d'éviter les erreurs d'interprétation des signaux sonores et pour protéger l'utilisateur contre des dangers imminent, une personne voyante l'a accompagnée pendant l'ensemble de ces déplacements.

Des scénarios de navigation différents avaient été sélectionnés et mise en œuvre afin d'éprouver la robustesse et de discerner les limites du système dans une variété d'espaces susceptibles d'être traversés lors de déplacements quotidiens par une personne malvoyante. Parmi ces environnements figuraient des allées piétonnes aux revêtements hétérogènes ou encore des quartiers résidentiels. Les parcours empruntés par le participant ont été enregistrés tout au long de l'expérience grâce aux données GPS émanant de l'antenne fixée sur le casque. Ces itinéraires ont ensuite été retranscrits sur des représentations cartographiques 2D de l'espace sur lequel les points rouge et vert représentent respectivement les points de départ et la destination désirée.

5.4.2 Résultat et interprétation

5.4.2.1 Suivi d'un trottoir et évitement d'obstacle

Les premières expérimentations se sont concentrées sur le suivi d'un itinéraire simple et relativement rectiligne, la principale difficulté résidant dans l'évitement d'obstacles statiques et la vérification que l'itinéraire proposé reste sur un chemin destiné aux piétons, évitant ainsi de guider l'utilisateur vers des zones dangereuses. En outre, la simplicité de cet espace permet d'évaluer la capacité de l'utilisateur à se déplacer en se fiant aux signaux sonores avant de s'aventurer dans des environnements plus complexes, nécessitant une capacité d'interprétation des signaux sonores plus fine.

La figure 5.11 illustre le premier parcours réalisé par le participant, soulignant les obstacles clés comme des barrières et des poteaux. L'utilisateur a dû naviguer autour de ces derniers pour parvenir à sa destination, réussissant à les identifier et à les contourner, bien que le passage des barrières disposées en échelonnement ait pris un temps considérable. Le reste du parcours, sans encombre significatif, a été franchi sans difficulté. De manière similaire, le second déplacement sur un trottoir le long d'une voie réservée aux véhicules, représenté à la figure 5.12, a démontré une capacité du système et de l'utilisateur à éviter les poteaux jalonnant le parcours et à différencier avec précision l'aire

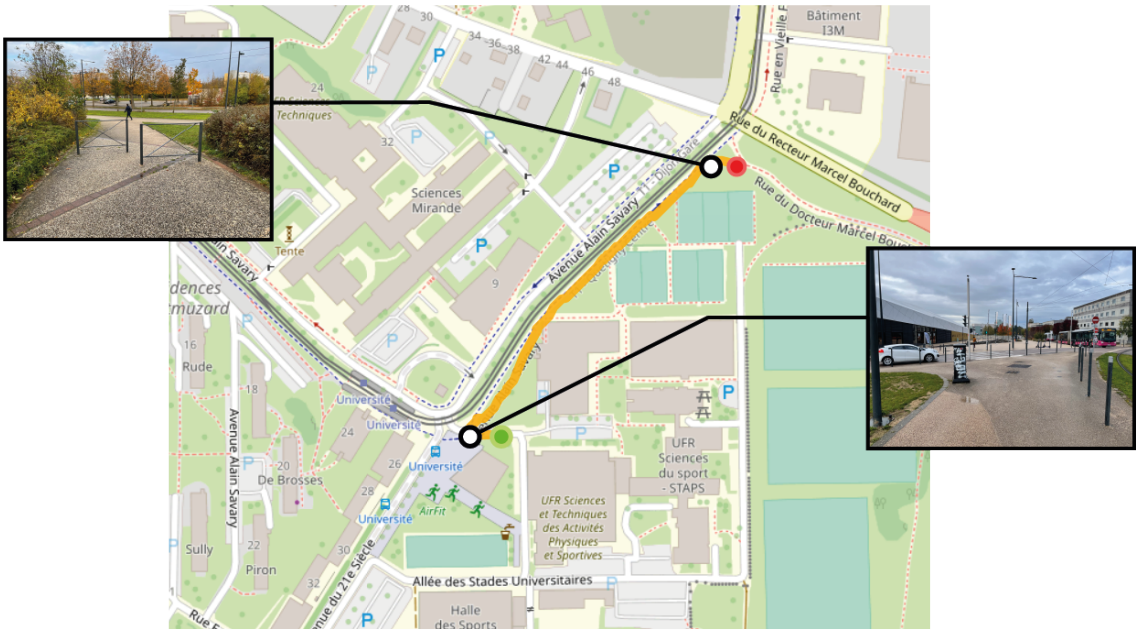


FIGURE 5.11 – Cartographie en 2D illustrant le parcours de l'utilisateur vers une destination indiquée par un point vert, partant d'une position marquée en rouge. Les points noirs représentent des obstacles statiques tels que des barrières et des poteaux.

piétonne de la route afin de garantir un itinéraire sécurisant.

FIGURE 5.12 – Cartographie en 2D illustrant le parcours de l'utilisateur sur un trottoir vers une destination indiquée par un point vert, partant d'une position marquée en rouge.

5.4.2.2 Navigation dans une zone piétonne

La seconde phase de l'expérimentation s'est déroulée sur un campus universitaire choisi explicitement pour son environnement piéton, peu dangereux et contrôlé, idéal pour évaluer l'interaction entre le participant et le système d'assistance. Malgré sa relative sécurité, le campus introduit des éléments dynamiques dans la tâche de navigation avec la présence de piétons ou de personnes circulant à vélo, augmentant la complexité du test par rapport aux précédentes expérimentations. De plus, le parcours prédéfini, menant à un bâtiment spécifique du campus, impliquait une navigation plus élaborée, traversant une multiplicité de sentiers et croisements. La carte présentée à la figure 5.13 représente l'itinéraire du participant dans cet espace de navigation. Les différentes phases propres à la méthode de navigation sont représentées dans ce parcours. En effet, le participant s'est d'abord dirigé dans la direction opposée au chemin naturel pour atteindre le nœud présent dans les données OpenStreetMap le plus proche de son point de départ. Après avoir atteint le point le plus proche, sa direction a changé pour suivre le chemin prédéfini, en utilisant des indices auditifs pour atteindre sa destination. En outre, le participant a évité les obstacles et les zones dangereuses, telles que la présence d'une pente importante

entre l'esplanade piétonne et la zone d'herbe. Le parcours long de 385 mètres et a été réalisé en 12 minutes, traduisant une allure moyenne d'environ 2 km à l'heure.

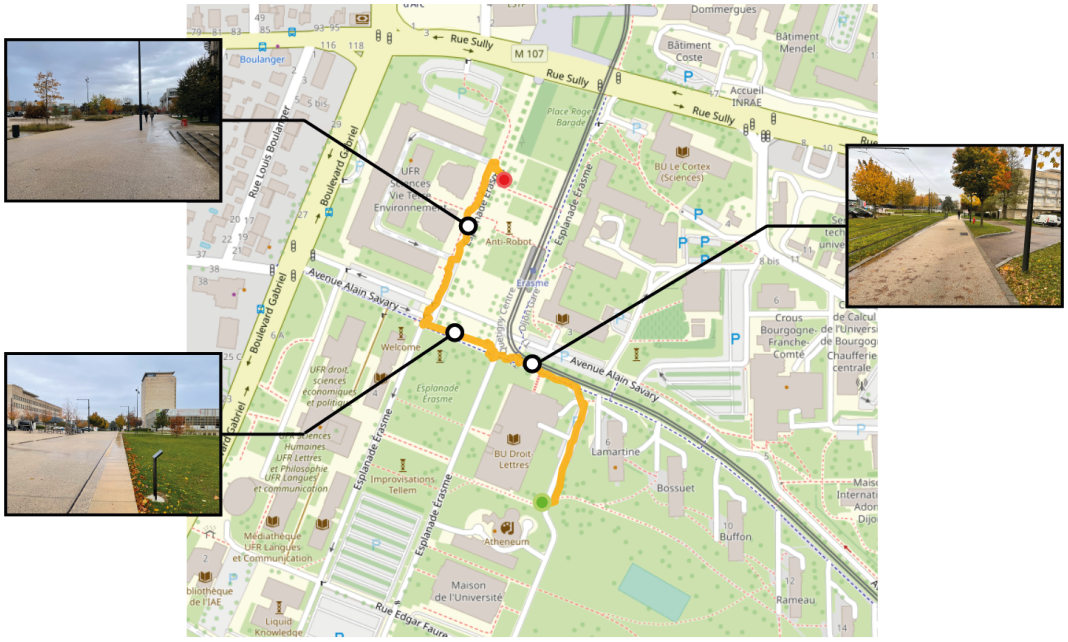


FIGURE 5.13 – Cartographie 2D illustrant le déplacement du participant pour atteindre la position marquée en vert à partir de la position rouge. Les points noirs mettent en évidence les points de repère importants le long du chemin, chacun associé à une information visuelle : la vue en haut à gauche montre une vaste zone de navigation, l'image en bas à gauche représente un espace avec une marche prononcée entre le chemin piétonnier et la zone herbeuse, et la vue la plus à gauche une intersection.

5.4.2.3 Navigation dans un espace partagé

L'espace précédant permettait d'élargir l'évaluation du système sur un espace plus complexe tout en conservant un espace contrôlé où les dangers étaient moindres. Il serait nécessaire de réitérer cette évaluation dans des espaces plus partagée entre les espaces piétons et réservée aux véhicules ainsi qu'intégrant une densité plus importante d'obstacles est nécessaire afin de visualiser des limites inhérentes à la méthode de navigation, de l'apport de vision artificielle ou d'erreur potentiel réalisée par le participant. Ainsi, la troisième partie de l'expérimentation s'est déroulée dans un environnement hybride caractérisé par l'intersection de voies piétonnes et de routes pour véhicules à moteur. Comme précédemment, le participant devait atteindre un bâtiment, mais cette fois-ci en traversant des intersections routières. Le tracé de son parcours, illustré à la figure 5.14 ; montre que l'utilisateur ne suit pas le trajet le plus court, mais donne la priorité aux chemins plus favorables aux piétons. Cette préférence a conduit le participant à

s'éloigner du trottoir et à emprunter un sentier longeant un bâtiment, avant de rejoindre la route principale. La figure montre également un détour par un chemin de terre moins visible (cadre en bas à gauche). Ce chemin, bien que moins visible, était présent dans les données cartographiques OpenStreetMap et a été identifié par l'algorithme de segmentation comme une voie accessible par un piéton. Le participant a parcouru l'ensemble du chemin d'environ 290 m en 10 minutes, ce qui correspond à un rythme moyen de 1,7 km par heure.

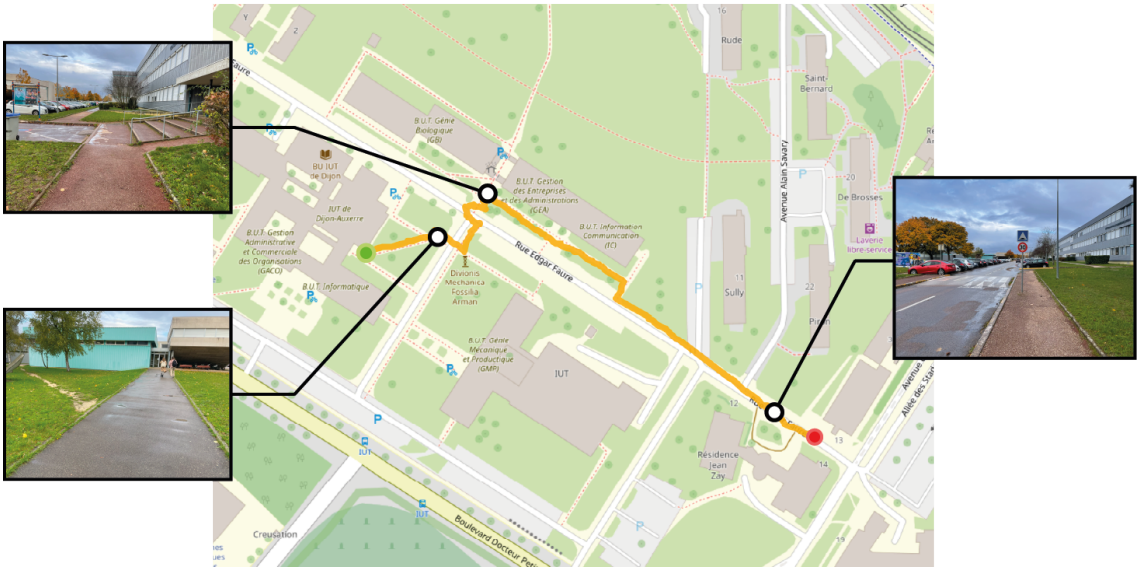


FIGURE 5.14 – Vue cartographique illustrant l'itinéraire du participant vers une destination. Les points noirs mettent en évidence les points de repère importants le long du trajet, chacun étant associé à une information visuelle : la vue en haut montrent la zone traversée sur la chaussée, et la vue positionnée en bas à gauche montre un sentier étroit emprunté par le piéton.

La dernière évaluation du dispositif a été effectuée dans un quartier résidentiel modérément dense, caractérisé par des trottoirs étroits, une circulation fluide et des véhicules garés le long des rues. Cet espace de navigation est proche des quartiers fréquentés régulièrement lors d'un déplacement dans une ville. Le participant a pu parcourir une distance de 232 mètres en 8 minutes sans aucun problème de navigation, hormis pour le passage de l'intersection entre les deux rues, sur laquelle une assistance de supervision fut nécessaire pour traverser. En effet, l'absence de marquage au sol pour les passages piétons à l'intersection a empêché le participant de traverser la route de manière autonome. Par conséquent, le superviseur l'a aidé à traverser la route une fois que l'absence d'obstacles avait été perçue par le participant. Néanmoins, il s'est habilement déplacé entre le trottoir et les voitures garées pendant l'exercice. L'itinéraire suivi est illustré à la figure 5.15, et l'intersection problématique est indiquée par le point noir à droite.

Au cours de l'ensemble de ces courtes expérimentations, nous avons évalué la capacité d'utilisateur dépourvu de connaissances visuelles sur son environnement et

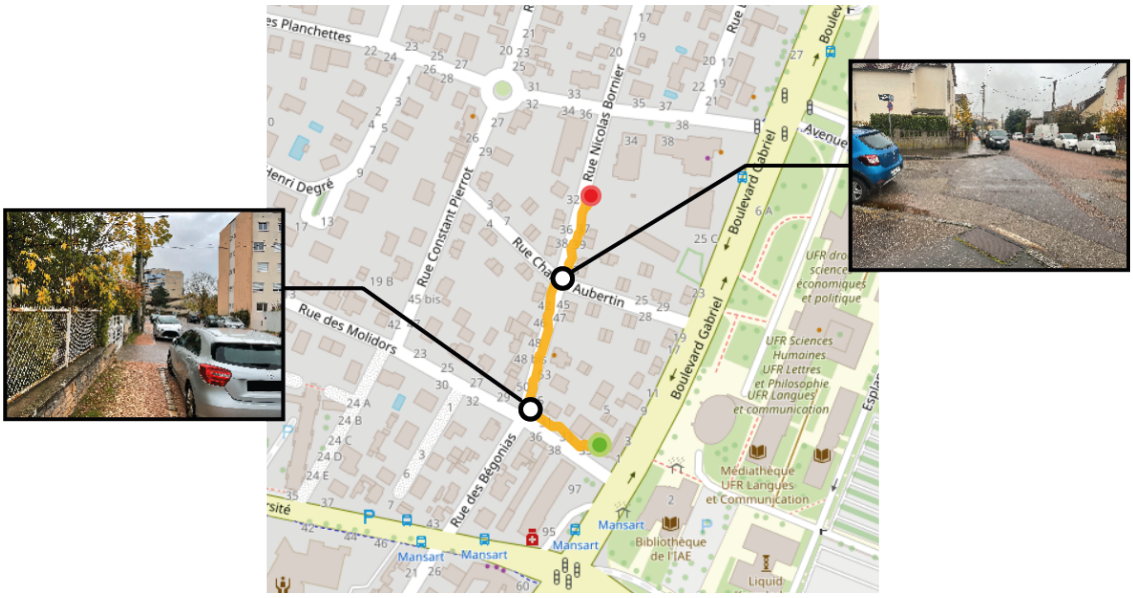


FIGURE 5.15 – Carte représentant le parcours d'un piéton dans un environnement urbain avec le point rouge et verte symbolisant le départ et la destination. Les deux points noirs mettent en évidence la zone de navigation, en particulier le trottoir et de l'intersection traversée.

équipée de notre système d'assistance à atteindre une suite de destinations demandées à l'aide de stimulus sonores spatialisées. Ces observations s'inscrivent dans la continuité de celles réalisées dans un bâtiment, où le participant a pu naviguer sans incident ni intrusion dans des zones à risque, malgré une certaine réticence née du manque de repères visuels habituels. La synergie des informations sonores relatives au parcours et aux obstacles environnants a contribué à pallier le déficit visuel, renforçant ainsi l'assurance et la sécurité lors de la navigation. Néanmoins, des tests supplémentaires dans des contextes plus complexes et variés ont révélé certaines limites. L'un des principaux freins était la dégradation des signaux GPS dans les paysages urbains plus étroits que les conditions d'évaluations proposées au participant. En effet, cette limitation est due à la présence de bâtiments réduisant le nombre de connexions satellitaire du GPS et par conséquent la précision de la géolocalisation. Malgré cela, l'intégration de données visuelles rectifiant la trajectoire a nettement accru la fiabilité du système, palliant en partie les déficiences du GPS. Cependant, l'approche proposée n'intègre pas de méthodologie propre pour gérer des passages sans marquage au sol ou pour reconnaître les signaux de feux piétons, bien que détecter par la méthode de détection d'obstacle nécessaire pour naviguer avec assurance dans des scénarios urbains complexes.

5.5 Conclusion

L'absence de connaissance visuelle et préalable sur un espace de navigation est un frein pour de nombreuses personnes malvoyantes pour se déplacer aisément sans assistance humaine. De surcroit, dans un environnement dangereux tel que peut l'être l'espace urbain, composé d'obstacles statiques et dynamiques ainsi que de zones inaccessibles aux piétons. Ces espaces peuvent inclure des trottoirs, des passages pour piétons dépourvus de surface podotactile, rendant la navigation difficile et risquée sans l'assistance visuelle ou physique adéquate. Face à cette contrainte réduisant fortement leurs autonomies, nous avons proposé au cours de ce chapitre un système d'assistance autonome à la navigation [179].

Ce dispositif offre un moyen d'atteindre un lieu désiré à partir de signaux sonores spatialisés en guidant à partir d'information géospatiale enregistrée ou acquise désignant un trajet adapté à son handicap. En effet, l'itinéraire proposé a été optimisé pour s'adapter aux complications éventuelles de certains axes routiers afin de permettre un déplacement agréable et sûre à partir d'extraction de données cartographiques. Cependant, bien que ces données soient cruciales dans la détermination d'une trajectoire vers une destination donnée, leur efficacité peut être compromise sans une prise en compte adéquate des menaces potentielles. Ainsi, une analyse visuelle de l'espace de navigation a été intégrée pour affiner le guidage de l'utilisateur vers des itinéraires sécurisés et pour identifier et signaler la présence d'obstacles, qu'ils soient immobiles ou en mouvement. L'efficacité et la pertinence de ces informations ont été validées lors d'expérimentations où les participants ont réussi à atteindre différentes destinations sans incidents, bien que certaines spécificités urbaines comme les intersections non régulées ou les feux de signalisation restreignent l'application généralisée de notre approche pour des contextes d'utilisation plus étendus. Ces performances de navigations sécurisées pour l'utilisateur ont été obtenues en tenant compte des exigences de robustesse et de réactivité cruciales pour les dispositifs d'assistance aux personnes malvoyantes, où toute détection erronée ou retardée des signaux peut avoir des conséquences graves. En effet, notre dispositif a été optimisé pour atteindre une faible latence et une consommation énergétique modérée à partir d'une cible matérielle embarquée de faible puissance et d'une architecture multi-processus réduisant le temps de traitement de chaque processus d'un système de substitution sensorielle.

Le prototype expérimental que nous avons développé constitue un pas préliminaire vers un système complet d'assistance à la navigation pour les individus malvoyants évoluant dans des milieux urbains. En effet, des limitations et des particularités inhérentes

aux espaces urbains inexploités durant ce chapitre empêchent encore une utilisation élargie du dispositif, mais posent les fondements d'une assistance aux déplacements plus autonome, sécurisante et confortable. Ce sentiment de confort pouvant être enrichi par une miniaturisation des dispositifs d'acquisition et de traitement avec une réduction de poids et d'encombrement.

6
Conclusion

Sommaire

6.1 Conclusion générale

L'assistance aux personnes malvoyantes sera un enjeu majeur et croissant lors des prochaines décennies. La perte d'autonomie associée à cette déficience, en particulier lors de leurs déplacements, impacte durement leurs vies sociales où chaque action fait face à de nombreux défis à surmonter. L'apport d'information complémentaire de manière autonome est fondamental, pour accroitre leur autonomie lors de la réalisation d'actions quotidiennes telles que la présence d'un obstacle sur la chaussée lors d'un déplacement. La thèse présentée s'est attachée au développement de méthodes d'assistance à destination des personnes aveugles à partir d'un mécanisme de substitution sensorielle d'informations visuelles en signaux sonores de courte durée. Le cœur de cette recherche s'est reposé sur un approfondissement des composantes propres à ces systèmes d'assistance en enrichissant les informations perçues par l'apport d'informations plus complexes et détaillées de son environnement. Ces avancées fondées sur un processus d'encodage de l'information visuo-spatial dans un espace auditif élaboré à partir d'indices auditifs naturellement employés chez une personne pour connaître l'origine d'une émission sonore, associée à des méthodes de vision artificielle pour déterminer la présence d'un élément pertinent. L'élaboration d'un encodage sonore à partir de facteurs indiciels inhérents au système auditif humain tels que l'intensité sonore ou la stéréophonie. L'encodage proposé exploite les capacités élevées de reconnaissance de l'origine d'une émission sonore au sein d'un environnement pour permettre une compréhension et une interprétation des stimulus sonores à la fois rapide, intuitive et précise.

L'encodage proposé a montré une capacité à percevoir la position spatiale d'un stimulus artificiel aisément lors d'une phase d'évaluation conduite lors de cette thèse [152], [153]. Ces capacités ont offert accès à une interprétation d'une information spatiale précise pouvant assister une personne lors d'un déplacement afin d'atteindre un lieu non familier à partir de l'émission sonore de la direction à suivre, mais également des obstacles à proximité. La fusion simultanée de ces informations assure une meilleure compréhension et interaction avec l'environnement, facilitant la navigation. Ce couplage d'informations dans un signal unique, enrichi par divers indices acoustiques, a été expérimenté dans le cadre d'une preuve de concept d'assistance à la navigation dans un bâtiment. Cette preuve de concept renseignant simultanément la direction à suivre pour atteindre la destination et les positions des éléments proches a permis de valider empiriquement la capacité d'une personne privée de faculté visuelle à se déplacer de manière autonome et sécurisée dans un milieu inconnu. En effet, au cours de l'expérimentation, le participant a pu rejoindre

un point spécifique tout en évitant des obstacles disposés sur son parcours, et ce, sans nécessiter de longues phases de familiarisation avec le dispositif [161]. Cette preuve de concept nous a permis de démontrer l'intérêt d'une telle approche en situation réelle et a ouvert la voie à l'adaptation du système pour des environnements de navigation encore plus exigeants, notamment dans les milieux urbains.

L'adaptation de ce système pour naviguer dans des espaces urbains, avec leurs défis spécifiques tels que la présence d'éléments dangereux, s'appuie sur des approches de vision artificielle agissant comme un substitut au système visuel, fournissant une perception et une interprétation de l'environnement de l'utilisateur. L'information recueillie par une caméra est analysée pour identifier le placement d'obstacles, qu'ils soient statiques ou dynamiques, ainsi que les zones à éviter. Ces processus fondés sur des techniques d'intelligence artificielle, notamment des réseaux de convolutions profonds, offrent une détermination précise tout en maintenant un faible temps d'inférence, assurant ainsi que l'utilisateur malvoyant dispose de suffisamment de temps pour interpréter les signaux sonores. Ces approches par apprentissage profond aussi bien de détection d'obstacle que de segmentation sémantique pour discriminer le sol ont été entraînées sur de larges bases d'apprentissage pour favoriser une robustesse des prédictions face aux différents environnements structurant un espace piéton. Dans ce cadre, nous avons créé une nouvelle base de donnée spécifique, axée spécifiquement sur la perspective d'un piéton, venant compléter celles existantes avec de nouveaux milieux d'acquisition. Ce nouveau jeu de données composé de diverses séquences vidéo de déplacements urbains piétonniers, où les éléments les plus fréquents dans de tels espaces ont été minutieusement annotés [162]. De surcroît, l'identification de zones sûres via l'analyse sémantique des scènes visuelles combinée à une méthode de planification de la trajectoire adaptée a permis de guider les utilisateurs sur des itinéraires.

La détermination de la trajectoire à partir de l'exploitation d'informations cartographiques détaillée pour définir une trajectoire confortable a été proposée afin d'exclure les zones complexes ou manquantes d'aménagements nécessaires, réduisant la charge cognitive des utilisateurs. Cette approche a été intégrée dans un système embarqué autonome, optimisé par la parallélisation des tâches entre différents cœurs de processeurs et un processeur graphique. Cette configuration nous a permis d'atteindre une performance optimale avec un fonctionnement temps-réelle combinée avec efficacité énergétique par son implantation adaptée aux dispositifs avec une puissance modérée [179]. Cependant, bien qu'une expérimentation de ce système ait démontré une aptitude à se déplacer librement uniquement à l'aide du système dans un espace urbain lors de différents scénarios d'évaluation, des limitations subsistent encore telles que le franchissement d'une intersec-

tion dépourvue de passage piéton. De surcroit, l'encombrement et la faible ergonomie du système d'acquisition peut provoquer une fatigue lors de longs déplacements.

Ces limitations invitent à affiner et à améliorer les méthodes d'assistance à la mobilité urbaine évoquées dans cette thèse. Des contextes spécifiques des espaces urbains restent à explorer, nécessitant le développement de méthodologies adaptées comme la présence de foule, d'intersection ou de travaux entravant la mobilité. En plus, envisager une optimisation de l'ergonomie du système, par sa miniaturisation ou le transfert de certaines fonctions à un dispositif externe, pourrait améliorer l'expérience d'un utilisateur malvoyant. L'implantation de ces approches de substitutions sensorielles sur un chien robotique permettrait d'alléger la charge portée par l'utilisateur tout en offrant des nouvelles perspectives grâce à une perception décentrée et de nouvelles formes d'interactions avec le système par l'apport de connaissances plus riches. L'intégration de retours haptiques, tels que la tension d'une laisse, avec des signaux sonores, donne des possibilités d'une duplicité de l'émission de l'information afin de favoriser leurs compréhensions ou bien de définir, de transmettre des informations de nature différentes par différentes modalités sensorielle, mais pourrait surcharger la capacité cognitive de l'utilisateur. Par ailleurs, l'interaction avec un système de chatbot, enrichie par des techniques avancées de vision artificielle pour la description de scènes, pourrait significativement augmenter la capacité d'une personne aveugle à percevoir et à comprendre son environnement selon les besoins spécifiques de l'utilisateur. L'ensemble de ces aspects fondés sur notre concept de substitution sensorielle visuo-sonore ouvre les perceptives du système assistance efficace et complet pour les personnes malvoyantes visant à diminuer significativement les freins à leur autonomie.

6.2 Publications

Articles de journaux internationaux

[1] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « uB-VisioGeoloc: An Image Sequences Dataset of Pedestrian Navigation Including Geolocalised-Inertial Information and Spatial Sound Rendering of the Urban Environment's Obstacles. », *Data In Brief (accepté)*, p. 1-10, 2023

[2] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « Outdoor Navigation Assistive System Based on Robust and Real-Time Visual–Auditory Substitution Approach », en, *Sensors*, t. 24, n^{o} 1, p. 166, 2023

[3] C. BORDEAU, F. SCALVINI, C. MIGNIOT, J. DUBOIS et M. AMBARD, « Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution », en, *Frontiers in Psychology*, t. 14, p. 1 079 998, 2023

Conférences internationales

[4] C. BORDEAU, F. SCALVINI, C. MIGNIOT, J. DUBOIS et M. AMBARD, « Distance perception of objects using visual-to-auditory sensory substitution: comparison of conversion methods based on sound intensity and envelope modulation », in *Auditory Perception, Cognition, & Action Meeting (APCAM)*, Boston, MA, 2022, p. 1

[5] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT, S. ARGON et J. DUBOIS, « Visual-auditory substitution device for indoor navigation based on fast visual marker detection », en, in *Signal-Image Technology & Internet-Based Systems (SITIS)*, Dijon, France : IEEE, 2022, p. 259-266

[6] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « Low-Latency Human-Computer Auditory Interface Based on Real-Time Vision Analysis », en, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore : IEEE, 2022, p. 36-40

Conférence nationale

[7] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « Système d'assistance à la mobilité en milieu urbain des personnes malvoyantes via une substitution de l'information visuelle par un signal auditif », in *Compas*, t. Poster, Annecy, France, 2023

7
ANNEXE

7.1 Calcul de distance

Dans l'apprentissage automatique, les mesures de distance sont essentielles pour divers algorithmes afin de mesurer la similarité ou la dissimilarité entre les points de données. Voici comment certaines mesures de distance sont utilisées dans différents contextes d'apprentissage automatique :

Distance de Manhattan ∼ Norme L1 : Distance entre deux points, calculée en ne considérant que les déplacements verticaux et horizontaux, similaire aux trajets dans des rues de Manhattan. Cette distance est déterminée par la somme des différences absolues entre les coordonnées respectives des points considérés.

$$\mathcal{L}1 = \sum(|x_i - y_i|) \tag{7.1}$$

Distance Euclidienne ∼ Norme L2 : Distance mesurée dans un espace euclidien, caractérisée par la racine carrée de l'addition des carrés des différences entre les coordonnées correspondantes des points.

$$\mathcal{L}2(x, y) = \sqrt{\sum(x_i - y_i)^2} \tag{7.2}$$

Distance de Tchebychev ∼ Norme L-infini : Cette distance est définie comme la plus grande différence absolue entre les composantes correspondantes de deux points. Elle est mathématiquement exprimée par :

$$\mathcal{L}_\infty = max(|x_i - y_i|) \tag{7.3}$$

Distance de Minkowski : Généralisation des distances de Manhattan et euclidienne. La distance de Minkowski est définie comme la racine p^{ime} de la somme des différences des coordonnées élevées à la puissance p. Les cas particuliers de la distance de Manhattan et euclidienne correspondent respectivement à $p = 1$ et $p = 2$.

$$\mathcal{M}(x, y) = \left(\sum_i |x_i - y_i|^p \right)^{1/p} \tag{7.4}$$

7.2 Fonction de perte

La sélection de la fonction de perte est essentielle et dépend directement des objectifs du modèle et de la nature des problèmes à résoudre, qu'il s'agisse de régression ou de classification. La fonction d'erreur employée pour un problème de régression est principalement l'erreur quadratique moyenne afin de quantifier l'écart entre les prédictions du modèle et les véritables valeurs. Dans le contexte de la classification, l'entropie croisée binaire est préférée pour les cas binaires, tandis que l'entropie croisée catégorielle est privilégiée pour les scénarios multiclasses ou multiétiquettes. Néanmoins, il existe d'autres fonctions de perte adaptées à des situations ou des exigences spécifiques.

Erreur quadratique moyenne ~ Mean Square Error : Métrique de régressions qui mesure l'amplitude moyenne des erreurs dans un groupe de prédiction sans considération de leurs directions. D'un point de vue mathématique, il s'agit d'une moyenne des différences au carré entre les prédictions et les valeurs espérées où l'ensemble des individus ont la même importance. Dans son expression mathématique symbolisée par l'équation 7.5, M représente le nombre d'échantillons de l'ensemble. Par ailleurs, y_i représente la prédiction du modèle pour un échantillon particulier x_i, alors que y_i indique la valeur réelle associée à l'échantillon.

$$\mathcal{L}(\hat{y}, y_i) = \frac{1}{M} \sum_{i=0}^{M} (y_i - \hat{y}_i)^2 \tag{7.5}$$

Entropie Croisée Binaire ~ Binary Cross Entropy (BCE) : Mesure de classification binaire de quantification de la divergence entre les probabilités prédites pour un ensemble de donnée d'entrée $\mathcal{X}$, et les véritables étiquettes. Une faible valeur de cette métrique indique que les prédictions du modèle sont en adéquation avec les vraies étiquettes. À l'inverse, une prédiction erronée est fortement sanctionnée en raison de la nature exponentielle de la fonction logarithmique employée.

$$\mathcal{L}(\hat{y}, y_i) = -\sum_{c=0}^{C=2}\sum_{j=1}^{M} y_{ij} \times \log(\hat{y}_{ij}) = -\frac{1}{M}\sum_{i=0}^{M} y_i \log(\hat{y}) - (1-y_i)\log(1-\hat{y}_i) \quad (7.6)$$

Entropie Croisée Catégorielle ∼ Categorical cross entropy : Extension de la BCE pour des problèmes de classification impliquant plus de deux classes. Elle mesure l'écart entre les probabilités prédites pour chaque classe et les véritables étiquettes one-hot encodées.

$$\mathcal{L}(\hat{y}, y_i) = -\sum_{c=0}^{C}\sum_{j=1}^{M} y_{ij} \times \log(\hat{y}_{ij}) \quad (7.7)$$

7.3 Métrique de classification et détection

7.3.1 Métrique de classification

L'évaluation de l'efficacité d'un modèle de classification repose sur des métriques précises qui reflètent sa capacité à faire des prédictions correctes et à distinguer clairement entre différentes classes. Ces métriques offrent une représentation quantitative du taux de succès du modèle et des erreurs potentielles qu'il pourrait commettre en classant des observations. Cette évaluation se base principalement sur les valeurs de matrice de confusion 7.1, qui catégorise les résultats en quatre classifications possibles, décrites ci-dessous :

- Vrai positif ∼ VP : Nombre ou le pourcentage d'observations réellement positives qui ont été correctement classées.
- Faux positif ∼ FP : Nombre ou le pourcentage d'observations réellement négatives qui ont été incorrectement classées.
- Vrai positif ∼ FN : Nombre ou le pourcentage d'observations réellement négatives qui ont été incorrectement classées.
- Faux positif ∼ VN : Nombre ou au pourcentage d'observations réellement négatives qui ont été correctement classées.

À partir des valeurs de la matrice de confusions, des métriques quantifiant les taux de précisions et d'erreurs peuvent être définies :

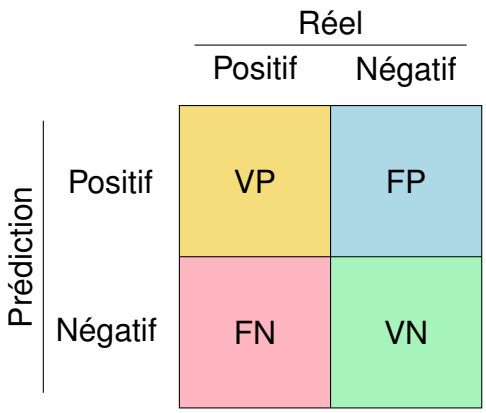


FIGURE 7.1 – Matrice de confusion.

Justesse ∼ Accuracy : Proportion d'estimations correctes par rapport au nombre total de prédictions effectuées. Dans le cas d'un jeu de données non équilibré, où la représentation des individus par classe n'est pas proportionnelle, la justesse ne reflète pas la performance réelle du modèle.

$$Accuracy = \frac{VP + VP}{VP + FP} \tag{7.8}$$

Précision : Proportion de nombre cas correctement classifié comme positif par rapport au nombre d'individus classifiés comme correct. Si

$$Precision = \frac{VP}{VP + FP} \tag{7.9}$$

Rappel ou sensibilité ∼ Recall : : Proportion de nombre cas correctement classifié comme positif par rapport au nombre d'individus positifs. Le rappel est souvent utilisé en tandem avec la précision, car cette dernière intègre la notion de faux positifs dans son calcul.

$$Rappel = \frac{VP}{VP + FN} \tag{7.10}$$

Score F1 : Métrique pour évaluer la performance conjointe de la précision et du rappel. Un score proche de 1 indique un bon équilibre entre la précision et le rappel tandis qu'une valeur faible traduit une faiblesse du rappel ou bien de la précision.

$$F1 = 2 \times \frac{precision \times rappel}{precision + rappel} \tag{7.11}$$

7.3.2 Métrique de détection d'objet

Le vecteur de sortie, combinant l'indice de classification et les coordonnées de la boîte sur l'image 2D, est mesurée par rapport aux valeurs réelles à partir de méthodes identiques aux méthodes de classification pour l'identifiant de classe et une métrique nommée intersection sur union pour la boîte englobante recherchant le taux de recouvrement. L'évaluation de l'IoU permet de s'assurer que le modèle détecte non seulement l'objet correctement, mais aussi qu'il le localise avec précision dans l'espace.

Intersection sur union ∼ IoU - Intersection over Union : Évalue l'exactitude de la localisation de l'objet en mesurant le degré de superposition entre la boîte englobante prédite par le modèle et la vérité terrain. Mathématiquement, elle est définie comme le ratio de l'aire de l'intersection des deux boîtes sur l'aire de leur union. Une IoU égale à 1 signifie que la boîte prédite coïncide parfaitement avec la boîte réelle, indiquant une localisation impeccable. Inversement, un IoU proche de 0 n'indique aucune superposition, traduisant une prédiction inexacte.

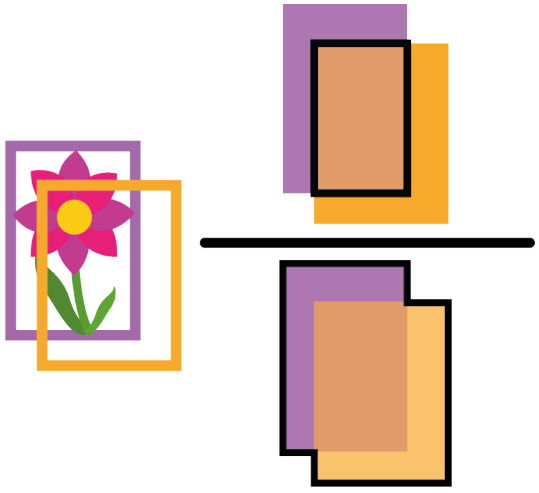


FIGURE 7.2 – Vue schématique du taux de recouvrement d'une prédiction en orange par rapport à la vérité terrain en violette.

7.3.3 Métrique de segmentation sémantique

Les métriques pour des taches de segmentation sémantiques regroupent principalement des métriques employées pour des taches de classifications et détection d'objet. Cependant, des métriques spécifiques sont également employées pour définir le degré

d'exactitude ou de similarité entre une prédiction et la vérité terrain.

Précision pixel : Vise à mesurer le pourcentage de pixels correctement classifiés par rapport au total des pixels dans l'image afin d'évaluer la concordance entre la segmentation prédite avec la vérité terrain pixel par pixel.

$$\text{Precision pixel} = \frac{\text{Nombre de pixels correctement classifiés}}{\text{Nombre total de pixels dans l'image}} \tag{7.12}$$

Dice : Évaluation de la similarité spatiale entre la vérité terrain et une prédiction. Similairement avec le score F1, une valeur proche de 1 indique une grande similarité, tandis qu'une valeur proche de 0 indique peu de similarité.

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{7.13}$$

Où X représente l'ensemble de pixels de la vérité terrain et Y représente l'ensemble de pixels de la prédiction.

7.4 Processus d'élaboration d'une méthode par d'apprentissage profond

L'élaboration et le déploiement d'un modèle basé sur les réseaux de neurones nécessitent une série d'étapes méthodiques afin de garantir son efficacité et sa fiabilité. Le processus global et subdivisé en huit sous parties distinctes, de la création d'un ensemble de donnée jusqu'à la mise en place du modèle au sein d'une application. Les différents processus représentés à la figure 7.3 sont les suivants :

1. **Préparation du jeu de données :** Avant d'entamer le processus d'entraînement, il est essentiel de préparer le jeu de donnée. Les données doivent être nettoyées pour éliminer tout élément indésirable, normalisées si besoin, puis divisées en sous-ensembles dédiés à l'entraînement, à la validation et aux tests.
2. **Augmentation des données :** Augmentation du nombre de données au sein de la base d'entraînement par des manipulations mathématiques ou géométriques des données existantes afin d'en créer de nouvelles artificiellement. Dans le domaine de la vision par ordinateur, l'augmentation est particulièrement cruciale étant donné la variété et la complexité des images dans le monde réel. Les techniques courantes comprennent : rotation, inversion, zoom, translation ou la modification des couleurs.

3. **Définition du modèle :** Une fois les données prêtes, la structure du réseau de neurones est conçue, déterminant ainsi l'architecture globale du modèle.
4. **Entraînement :** Durant cette phase, le modèle apprend en optimisant ses paramètres selon une méthode d'apprentissage spécifique.
5. **Validation :** Processus d'évaluation des capacités de généralisation du modèle sur un jeu de données distinct de celui de l'entraînement. Une divergence marquée des performances entre l'entraînement et la validation peut indiquer un sous-apprentissage ou un sur-apprentissage. Le sous-apprentissage (under-fitting) se manifeste par une mauvaise adaptation du modèle aux données, tandis que le sur-apprentissage (over-fitting) résulte d'un apprentissage trop spécifique aux données d'entraînement, rendant le modèle incapable de bien généraliser sur de nouvelles données.
6. **Optimisation du modèle :** Optimisation des hyperparamètres du réseau si le modèle n'atteint pas les performances souhaitées par un ajustement de certains hyperparamètre, d'introduction de régulation ou modification de l'architecture. De plus, une phase de quantification peut être mise en œuvre pour réduire la complexité du modèle, optimisant ainsi le temps d'exécution.
7. **Déploiement du modèle :** Cette étape consiste à intégrer le modèle, avec sa structure et ses paramètres finalisés, dans une application ou un système de production où des inférences sont réalisées sur des données inédites.

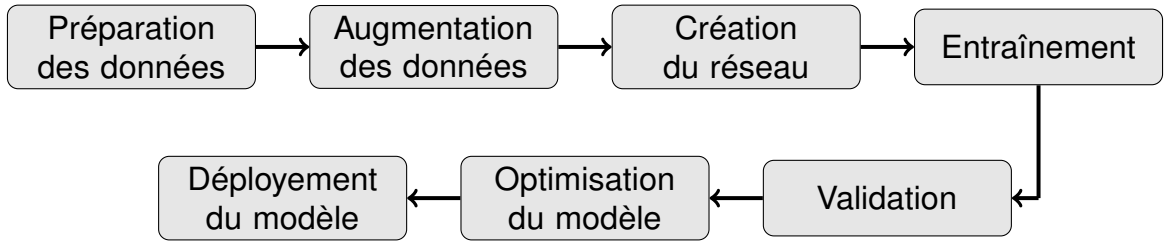


FIGURE 7.3 – Processus d'élaboration et de déploiement d'un réseau de neurones artificiels.

References

[1] J. D. STEINMETZ, R. R. A. BOURNE, P. S. BRIANT et al., « Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study », en, *The Lancet Global Health*, t. 9, n° 2, e144-e160, 2021.

[2] W. H. ORGANIZATION, « Rapport mondial sur la vision », fr, World Health Organization, Genève, rapp. tech., 2020, Section: xviii, 170 p., p. 1-192.

[3] L. DANDONA et R. DANDONA, « Revision of visual impairment definitions in the International Statistical Classification of Diseases », en, *BMC Medicine*, t. 4, n° 1, p. 1-7, 2006.

[4] R. BOURNE, J. D. STEINMETZ, S. FLAXMAN et al., « Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study », en, *The Lancet Global Health*, t. 9, n° 2, e130-e143, 2021, Number: 2.

[5] UNITED NATIONS, *World Population Prospects 2022: Summary of Results* (Statistical Papers - United Nations (Ser. A), Population and Vital Statistics Report), en. United Nations, 2022.

[6] THUCYDIDE, *LA GUERRE DU PELOPONNESE*, fr. Grèce.

[7] C. HUSQUIN, « Fiat Lux ! Cécité et déficiences visuelles à Rome : réalités et mythologies, des ténèbres à la lumière », fr, *Pallas*, n° 106, p. 243-256, 2018.

[8] Z. WEYGAND, « Les aveugles dans la société française: Représentations et institutions du Moyen-Âge au xixe siècle », fr, *Revue d'éthique et de théologie morale*, t. 256, n° HS, p. 65, 2009.

[9] G. W. IGUNE, « Inclusion of Blind Children In Primary Schools », en, rapp. tech., 2009, p. 1-108.

[10] J. JEREMIAH, « Satisfaction of students with visual impairment within different school settings », en, *Educational Specialist*, t. 10, p. 1-32, 2015.

[11] M. C. MCDONNALL et Z. SUI, « Employment and Unemployment Rates of People Who Are Blind or Visually Impaired: Estimates from Multiple Sources », en, *Journal of Visual Impairment & Blindness*, t. 113, n° 6, p. 481-492, 2019.

[12] H. HARRABI, M.-J. AUBIN, M. V. ZUNZUNEGUI, S. HADDAD et E. E. FREEMAN, « Visual Difficulty and Employment Status in the World », en, *PLoS ONE*, t. 9, n° 2, Y. ZHENG, éd., e88306, 2014.

[13] A. P. MARQUES, J. RAMKE, J. CAIRNS et al., « Global economic productivity losses from vision impairment and blindness », en, *eClinicalMedicine*, t. 35, p. 100 852, 2021.

[14] E. CROFT, « Experiences of Visually Impaired and Blind Students in UK Higher Education: An Exploration of Access and Participation », en, *Scandinavian Journal of Disability Research*, t. 22, n° 1, p. 382-392, 2020.

[15] ROOPCHUND RANDHIR, KHIRODHUR LATASHA, PANYANDEE TOORAIVEN et BAPPOO MONISHAN, « Analyzing the Impact of Sensory Marketing on Consumers: A Case Study of KFC », en, *Journal of US-China Public Administration*, t. 13, n° 4, p. 278-292, 2016.

[16] M. EIMER, « Multisensory Integration: How Visual Experience Shapes Spatial Perception », en, *Current Biology*, t. 14, n° 3, R115-R117, 2004.

[17] F. GAUNET, J.-L. MARTINEZ et C. THINUS-BLANC, « Early-Blind Subjects' Spatial Representation of Manipulatory Space: Exploratory Strategies and Reaction to Change », en, *Perception*, t. 26, nº 3, p. 345-366, 1997.

[18] L. TCHEANG, H. H. BÜLTHOFF et N. BURGESS, « Visual influence on path integration in darkness indicates a multimodal representation of large-scale space », en, *National Academy of Sciences*, t. 108, nº 3, p. 1152-1157, 2011, Number: 3.

[19] L. PUTZAR, I. GOERENDT, K. LANGE, F. RÖSLER et B. RÖDER, « Early visual deprivation impairs multisensory interactions in humans », en, *Nature Neuroscience*, t. 10, nº 10, p. 1243-1245, 2007.

[20] K. PETRINI, P. R. JONES, L. SMITH et M. NARDINI, « Hearing Where the Eyes See: Children Use an Irrelevant Visual Cue When Localizing Sounds », en, *Child Development*, t. 86, nº 5, p. 1449-1457, 2015, Number: 5.

[21] D. J. BROWN et M. J. PROULX, « Audio–Vision Substitution for Blind Individuals: Addressing Human Information Processing Capacity Limitations », en, *Journal of Selected Topics in Signal Processing*, t. 10, nº 5, p. 924-931, 2016.

[22] M. I. POSNER, « Orienting of Attention », en, *Quarterly Journal of Experimental Psychology*, t. 32, nº 1, p. 1-24, 1980.

[23] C. D. GILBERT et T. N. WIESEL, « Receptive field dynamics in adult primary visual cortex », en, *Nature*, t. 356, nº 6365, p. 150-152, 1992, Number: 6365.

[24] D. BAVELIER et H. J. NEVILLE, « Cross-modal plasticity: where and how? », en, *Nature Reviews Neuroscience*, t. 3, nº 6, p. 443-452, 2002.

[25] H. PETER R., « Synaptic density in human frontal cortex — Developmental changes and effects of aging », en, *Brain Research*, t. 163, nº 2, p. 195-205, 1979, Number: 2.

[26] L. G. COHEN, R. A. WEEKS, N. SADATO, P. CELNIK, K. ISHII et M. HALLETT, « Period of susceptibility for cross-modal plasticity in the blind », en, *Annals of Neurology*, t. 45, nº 4, p. 451-460, 1999.

[27] N. SADATO, T. OKADA, M. HONDA et Y. YONEKURA, « Critical Period for Cross-Modal Plasticity in Blind Humans: A Functional MRI Study », en, *NeuroImage*, t. 16, nº 2, p. 389-400, 2002.

[28] L. B. MERABET et A. PASCUAL-LEONE, « Neural reorganization following sensory loss: the opportunity of change », en, *Nature Reviews Neuroscience*, t. 11, nº 1, p. 44-52, 2010.

[29] L. WHITMARSH, « The Benefits of Guide Dog Ownership », en, *Visual Impairment Research*, t. 7, nº 1, p. 27-42, 2005.

[30] J. M. RICKLY, N. HALPERN, M. HANSEN et J. WELSMAN, « Travelling with a Guide Dog: Experiences of People with Vision Impairment », en, *Sustainability*, t. 13, nº 5, p. 2840, 2021.

[31] M. ORMEROD, R. NEWTON, H. MACLENNAN et al., « Older people's experiences of using tactile paving », en, *Proceedings of the Institution of Civil Engineers - Municipal Engineer*, t. 168, nº 1, p. 3-10, 2015.

[32] J. FERNÁNDEZ GONZÁLEZ et A. GONGAL, « Unidirectional Tactile Paving: Circulation for the Visually Impaired », en, in *Studies in Health Technology and Informatics*, I. GAROFOLO, G. BENCINI et A. ARENGHI, éd., IOS Press, 2022.

[33] W. ELMANNAI et K. ELLEITHY, « Sensor-Based Assistive Devices for Visually-Impaired People: Current Status, Challenges, and Future Directions », en, *Sensors*, t. 17, n° 3, p. 565, 2017.

[34] V. GAILLET, A. CUTRONE, F. ARTONI et al., « Spatially selective activation of the visual cortex via intraneural stimulation of the optic nerve », en, *Nature Biomedical Engineering*, t. 4, n° 2, p. 181-194, 2019.

[35] S. R. KANE, S. F. COGAN, J. EHRLICH, T. D. PLANTE et D. B. MCCREERY, « Electrical performance of penetrating microelectrodes chronically implanted in cat cortex », en, in *Medicine and Biology Society*, Boston, MA : IEEE, 2011, p. 5416-5419.

[36] K. STINGL, R. SCHIPPERT, K. U. BARTZ-SCHMIDT et al., « Interim Results of a Multicenter Trial with the New Electronic Subretinal Implant Alpha AMS in 15 Patients Blind from Inherited Retinal Degenerations », en, *Frontiers in Neuroscience*, t. 11, p. 445, 2017.

[37] L. DA CRUZ, J. D. DORN, M. S. HUMAYUN et al., « Five-Year Safety and Performance Results from the Argus II Retinal Prosthesis System Clinical Trial », en, *Ophthalmology*, t. 123, n° 10, p. 2248-2254, 2016.

[38] A. J. LOWERY, J. V. ROSENFELD, P. M. LEWIS et al., « Restoration of vision using wireless cortical implants: The Monash Vision Group project », en, in *Engineering in Medicine and Biology Society (EMBC)*, Milan : IEEE, 2015, p. 1041-1044.

[39] W. H. BOSKING, M. S. BEAUCHAMP et D. YOSHOR, « Electrical Stimulation of Visual Cortex: Relevance for the Development of Visual Cortical Prosthetics », en, *Annual Review of Vision Science*, t. 3, n° 1, p. 141-166, 2017.

[40] L. DA CRUZ, B. F. COLEY, J. DORN et al., « The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss », en, *British Journal of Ophthalmology*, t. 97, n° 5, p. 632-636, 2013.

[41] M. S. HUMAYUN, « Intraocular retinal prosthesis », en, *Transactions of the American Ophthalmological Society*, t. 99, p. 271-300, 2001.

[42] M. S. HUMAYUN, J. D. DORN, L. DA CRUZ et al., « Interim Results from the International Trial of Second Sight's Visual Prosthesis », en, *Ophthalmology*, t. 119, n° 4, p. 779-788, 2012.

[43] S. SZPIRO, Y. ZHAO et S. AZENKOT, « Finding a store, searching for a product: a study of daily challenges of low vision people », en, in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Heidelberg Germany : ACM, 2016, p. 61-72.

[44] L. STEARNS, L. FINDLATER et J. E. FROEHLICH, « Design of an Augmented Reality Magnification Aid for Low Vision Users », en, in *Conference on Computers and Accessibility (ASSETS)*, Galway Ireland : ACM, 2018, p. 28-39.

[45] Y. CAI, Y. WANG et P. B. FREEMAN, « Real-time imaging processing with augmented reality glasses for mobile low-vision users », en, *Electronic Imaging*, t. 35, n° 10, p. 1-5, 2023.

[46] P. BACH-Y-RITA et S. W. KERCEL, « Sensory substitution and the human–machine interface », en, *Trends in Cognitive Sciences*, t. 7, n° 12, p. 541-546, 2003.

[47] R. PASSINI, G. PROULX et C. RAINVILLE, « The Spatio-Cognitive Abilities of the Visually Impaired Population », en, *Environment and Behavior*, t. 22, n° 1, p. 91-118, 1990.

[48] W. R. WIENER, R. L. WELSH et B. B. BLASCH, « Foundations of Orientation and Mobility - 3rd Edition », *American Foundation for the Blind*, p. 854, 2010.

[49] N. SILBERMAN, D. HOIEM, P. KOHLI et R. FERGUS, « Indoor Segmentation and Support Inference from RGBD Images », en, in *European Conference on Computer Vision (ECCV)*, D. HUTCHISON, T. KANADE, J. KITTLER et al., éd., t. 7576, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg : Springer, 2012, p. 746-760.

[50] M. HERSH, « Route learning by blind and partially sighted people », *Journal of Blindness Innovation and Research*, t. 10, nº 2, p. 1-33, 2020.

[51] J. M. LOOMIS, R. L. KLATZKY et N. A. GIUDICE, « - Sensory Substitution of Vision: Importance of Perceptual and Cognitive Processing », en, in *Assistive Technology for Blindness and Low Vision*, 0e éd., CRC Press, 2018, p. 179-210.

[52] M. J. PROULX, J. GWINNUTT, S. DELL'ERBA, S. LEVY-TZEDEK, A. A. DE SOUSA et D. J. BROWN, « Other ways of seeing: From behavior to neural mechanisms in the online "visual" control of action with sensory substitution », en, *Restorative Neurology and Neuroscience*, t. 34, nº 1, p. 29-44, 2015.

[53] W. SCHNEIDER et R. M. SHIFFRIN, « Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention », en, *Psychological Review*, t. 84, nº 1, p. 1-66, 1977.

[54] Á. KRISTJÁNSSON, A. MOLDOVEANU, Ó. I. JÓHANNESSON et al., « Designing sensory-substitution devices: Principles, pitfalls and potential », en, *Restorative Neurology and Neuroscience*, t. 34, nº 5, p. 769-787, 2016.

[55] G. E. LEGGE, C. MADISON, B. N. VAUGHN, A. M. Y. CHEONG et J. C. MILLER, « Retention of high tactile acuity throughout the life span in blindness », en, *Perception & Psychophysics*, t. 70, nº 8, p. 1471-1488, 2008.

[56] A. R. GARCIA, R. FONSECA et A. DURAN, « Electronic long cane for locomotion improving on visual impaired people. A case study », en, in *Pan American Health Care Exchanges (PAHCE)*, Rio de Janeiro, Brazil : IEEE, 2011, p. 58-61.

[57] A. COSGUN, E. A. SISBOT et H. I. CHRISTENSEN, « Guidance for human navigation using a vibro-tactile belt interface and robot-like motion planning », en, in *International Conference on Robotics and Automation (ICRA)*, Hong Kong, China : IEEE, 2014, p. 6350-6355.

[58] A. COSGUN, E. A. SISBOT et H. I. CHRISTENSEN, « Evaluation of rotational and directional vibration patterns on a tactile belt for guiding visually impaired people », en, in *Haptics Symposium (HAPTICS)*, Houston, TX, USA : IEEE, 2014, p. 367-370.

[59] M. A. HELLER, « Active and passive tactile braille recognition », en, *Bulletin of the Psychonomic Society*, t. 24, nº 3, p. 201-202, 1986.

[60] J. F. OLIVEIRA, « The path force feedback belt », en, in *Conference on Information Technology in Asia (CITA)*, Kota Samarahan, Malaysia : IEEE, 2013, p. 1-6.

[61] T. H. NGUYEN, T. H. NGUYEN, T. L. LE, T. T. H. TRAN, N. VUILLERME et T. P. VUONG, « A wearable assistive device for the blind using tongue-placed electrotactile display: Design and verification », en, in *International Conference on Control, Automation and Information Sciences (ICCAIS)*, Nha Trang, Vietnam : IEEE, 2013, p. 42-47.

[62] P. BACH-Y-RITA, C. C. COLLINS, F. A. SAUNDERS, B. WHITE et L. SCADDEN, « Vision Substitution by Tactile Image Projection », en, *Nature*, t. 221, nº 5184, p. 963-964, 1969.

[63] A. NASSER, K.-N. KENG et K. ZHU, « ThermalCane: Exploring Thermotactile Directional Cues on Cane-Grip for Non-Visual Navigation », en, in *Conference on Computers and Accessibility*, Greece : ACM, 2020, p. 1-12.

[64] Z. SUN, M. ZHU, X. SHAN et C. LEE, « Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions », en, *Nature Communications*, t. 13, n° 1, p. 5224, 2022.

[65] M. ZHU, Z. SUN et C. LEE, « Soft Modular Glove with Multimodal Sensing and Augmented Haptic Feedback Enabled by Materials' Multifunctionalities », en, *ACS Nano*, t. 16, n° 9, p. 14 097-14 110, 2022.

[66] K. KACZMAREK et S. HAASE, « Pattern identification as a function of stimulation on a fingertip-scanned electrotactile display », en, *Neural Systems and Rehabilitation Engineering*, t. 11, n° 3, p. 269-275, 2003.

[67] D. AHMETOVIC, C. GLEASON, C. RUAN, K. KITANI, H. TAKAGI et C. ASAKAWA, « NavCog: a navigational cognitive assistant for the blind », en, in *Human-Computer Interaction with Mobile Devices and Services*, Florence Italy : ACM, 2016, p. 90-99.

[68] E. M. HAVIK, A. C. KOOIJMAN et F. J. J. M. STEYVERS, « The Effectiveness of Verbal Information Provided by Electronic Travel Aids for Visually Impaired Persons », en, *Journal of Visual Impairment & Blindness*, t. 105, n° 10, p. 624-637, 2011.

[69] S. S. STEVENS, J. VOLKMANN et E. B. NEWMAN, « A Scale for the Measurement of the Psychological Magnitude Pitch », en, *The Journal of the Acoustical Society of America*, t. 8, n° 3, p. 185-190, 1937.

[70] D. TOLLIN et T. YIN, « Sound Localization: Neural Mechanisms », en, in *Encyclopedia of Neuroscience*, Elsevier, 2009, p. 137-144.

[71] W. L. GULICK, G. A. GESCHEIDER et R. D. FRISINA, *Hearing: Physiological acoustics, neural coding, and psychoacoustics.* Oxford University Press, 1989.

[72] T. T. SANDEL, D. C. TEAS, W. E. FEDDERSEN et L. A. JEFFRESS, « Localization of Sound from Single and Paired Sources », en, *The Journal of the Acoustical Society of America*, t. 27, n° 5, p. 842-852, 1955.

[73] T. R. LETOWSKI et S. T. LETOWSKI, « Auditory Spatial Perception: Auditory Localization: » en, Defense Technical Information Center, Fort Belvoir, VA, rapp. tech., 2012.

[74] N. KOLOTZEK, G. GOMEZ et B. U. SEEBER, « The effect of head turning on sound localization with hearing-aid satellites », en, rapp. tech., 2017, p. 1-2.

[75] S. K. ROFFLER et R. A. BUTLER, « Localization of Tonal Stimuli in the Vertical Plane », en, *The Journal of the Acoustical Society of America*, t. 43, n° 6, p. 1260-1266, 1968.

[76] J. C. MIDDLEBROOKS et D. M. GREEN, « Sound Localization by Human Listeners », en, *Annual Review of Psychology*, t. 42, n° 1, p. 135-159, 1991.

[77] J. HEBRANK et D. WRIGHT, « Spectral cues used in the localization of sound sources on the median plane », en, *The Journal of the Acoustical Society of America*, t. 56, n° 6, p. 1829-1834, 1974.

[78] G. ANDÉOL, E. A. MACPHERSON et A. T. SABIN, « Sound localization in noise and sensitivity to spectral shape », en, *Hearing Research*, t. 304, p. 20-27, 2013.

[79] J. BLAUERT et R. A. BUTLER, « *Spatial Hearing: The Psychophysics of Human Sound Localization* by Jens Blauert », en, *The Journal of the Acoustical Society of America*, t. 77, n° 1, p. 334-335, 1985.

[80] P. MEIJER, « An experimental system for auditory image representations », en, *Biomedical Engineering*, t. 39, n° 2, p. 112-121, 1992.

[81] E. STRIEM-AMIT, M. GUENDELMAN et A. AMEDI, « 'Visual' Acuity of the Congenitally Blind Using Visual-to-Auditory Sensory Substitution », en, *PLoS ONE*, t. 7, n° 3, A. SERINO, éd., e33136, 2012.

[82] J. WARD et P. MEIJER, « Visual experiences in the blind induced by an auditory sensory substitution device », en, *Consciousness and Cognition*, t. 19, n° 1, p. 492-500, 2010.

[83] M. AUVRAY, S. HANNETON et J. K. O'REGAN, « Learning to Perceive with a Visuo — Auditory Substitution System: Localisation and Object Recognition with 'The Voice' », en, *Perception*, t. 36, n° 3, p. 416-430, 2007.

[84] S. ABBOUD, S. HANASSY, S. LEVY-TZEDEK, S. MAIDENBAUM et A. AMEDI, « EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution », en, *Restorative Neurology and Neuroscience*, t. 32, n° 2, p. 247-257, 2014.

[85] S. LEVY-TZEDEK, S. HANASSY, S. ABBOUD, S. MAIDENBAUM et A. AMEDI, « Fast, accurate reaching movements with a visual-to-auditory sensory substitution device », en, *Restorative Neurology and Neuroscience*, t. 30, n° 4, p. 313-323, 2012.

[86] A. ALFARO, Á. BERNABEU, C. AGULLÓ, J. PARRA et E. FERNÁNDEZ, « Hearing colors: an example of brain plasticity », en, *Frontiers in Systems Neuroscience*, t. 9, p. 1-9, 2015.

[87] J. D. GOMEZ VALENCIA, « A computer-vision based sensory substitution device for the visually impaired (See ColOr) », en, thèse de doct., Université de Genève, 2014.

[88] G. BOLOGNA, B. DEVILLE et T. PUN, « Blind Navigation along a Sinuous Path by Means of the See ColOr Interface », en, in *Bioinspired Applications in Artificial and Natural Computation*, t. 5602, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, p. 235-243.

[89] C. CAPELLE, C. TRULLEMANS, P. ARNO et C. VERAART, « A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution », en, *Biomedical Engineering*, t. 45, n° 10, p. 1279-1293, 1998.

[90] S. HANNETON, M. AUVRAY et B. DURETTE, « The Vibe: a versatile vision-to-audition sensory substitution device », en, *Applied Bionics and Biomechanics*, t. 7, n° 4, p. 269-276, 2010.

[91] B. DURETTE, N. LOUVETON, D. ALLEYSSON et J. HÉRAULT, « Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE », en, in *Workshop on Computer Vision Applications for the Visually Impaired*, Marseille, France, 2008, p. 1-14.

[92] A. MHAISH, T. GHOLAMALIZADEH et D. J. DUFF, « Assessment of a Visual to Spatial-Audio Sensory Substitution System », en, in *Signal Processing and Communication Application Conference*, Zonguldak, Turkey : IEEE, 2016, p. 245-248.

[93] MIN NIE, JIE REN, ZHENGJUN LI et al., « SoundView: An auditory guidance system based on environment understanding for the visually impaired people », en, in *Engineering in Medicine and Biology Society*, Minneapolis, MN : IEEE, 2009, p. 7240-7243.

[94] V. PLANINEC, J. REIJNIERS, M. HORVAT, H. PEREMANS et K. JAMBROŠIĆ, « The Accuracy of Dynamic Sound Source Localization and Recognition Ability of Individual Head-Related Transfer Functions in Binaural Audio Systems with Head Tracking », en, *Applied Sciences*, t. 13, n° 9, p. 5254, 2023.

[95] L. COMMÈRE, S. U. N. WOOD et J. ROUAT, *Evaluation of a Vision-to-Audition Substitution System that Provides 2D WHERE Information and Fast User Learning*, en, arXiv:2010.09041 [cs], 2020.

[96] S. FERRAND, F. ALOUGES et M. AUSSAL, « An Augmented Reality Audio Device Helping Blind People Navigation », en, in *Computers Helping People with Special Needs*, K. MIESENBERGER et G. KOUROUPETROGLOU, éd., t. 10897, Series Title: Lecture Notes in Computer Science, Cham : Springer International Publishing, 2018, p. 28-35.

[97] F. RIBEIRO, D. FLORENCIO, P. A. CHOU et Z. ZHANG, « Auditory augmented reality: Object sonification for the visually impaired », en, in *Workshop on Multimedia Signal Processing (MMSP)*, Banff, AB, Canada : IEEE, 2012, p. 319-324.

[98] D. AGUERREVERE, « Portable 3D Sound / Sonar Navigation System for Blind Individuals », en, in *Latin American and Caribbean Conference for Engineering and Technology*, Miami, FL, USA, 2004, p. 1-6.

[99] J. J. LÓPEZ et A. GONZÁLEZ, « 3-D Audio with Video Tracking for Multimedia Environments », en, *Journal of New Music Research*, t. 30, n° 3, p. 271-277, 2001.

[100] A. HABIB, M. ISLAM, M. KABIR, M. MREDUL et M. HASAN, « Staircase Detection to Guide Visually Impaired People: A Hybrid Approach », en, *Revue d'Intelligence Artificielle*, t. 33, n° 5, p. 327-334, 2019.

[101] W. K. MUTLAG, S. K. ALI, Z. M. AYDAM et B. H. TAHER, « Feature Extraction Methods: A Review », en, *Journal of Physics: Conference Series*, t. 1591, n° 1, p. 012 028, 2020.

[102] M. W. NASRUDIN, N. S. YAAKOB, N. A. ABDUL RAHIM, M. Z. ZAHIR AHMAD, N. RAMLI et M. S. AZIZ RASHID, « Moment Invariants Technique for Image Analysis and Its Applications: A Review », en, *Journal of Physics: Conference Series*, t. 1962, n° 1, p. 012 028, 2021.

[103] N. DALAL et B. TRIGGS, « Histograms of Oriented Gradients for Human Detection », en, in *Computer Vision and Pattern Recognition (CVPR)*, t. 1, San Diego, CA, USA : IEEE, 2005, p. 886-893.

[104] T. LINDEBERG, « Scale Invariant Feature Transform », en, *Scholarpedia*, t. 7, n° 5, p. 10 491, 2012.

[105] A. D. GORDON, L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN et C. J. STONE, « Classification and Regression Trees. », en, *Biometrics*, t. 40, n° 3, p. 874, 1984.

[106] MCCULLOCH, « A logical calculus of the ideas immanent in nervous activity », en, *Bulletin of Mathematical Biophysic*, p. 115-133, 1943.

[107] F. ROSENBLATT, « The perceptron: A probabilistic model for information storage and organization in the brain. », en, *Psychological Review*, t. 65, n° 6, p. 386-408, 1958.

[108] X. CHEN, S. XIANG, L. L. CHENG et C.-H. PAN, « Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks », en, *Geoscience and Remote Sensing Letters*, t. 11, n° 10, p. 1797-1801, 2014.

[109] R. GIRSHICK, J. DONAHUE, T. DARRELL et J. MALIK, « Rich feature hierarchies for accurate object detection and semantic segmentation », en, Columbus, OH, USA : IEEE, 2014, p. 580-587.

[110] R. GIRSHICK, « Fast R-CNN », en, Santiago, Chile : IEEE, 2015, p. 1440-1448.

[111] S. REN, K. HE, R. GIRSHICK et J. SUN, « Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks », en, *Transactions on Pattern Analysis and Machine Intelligence*, t. 39, n° 6, p. 1137-1149, 2017.

[112] J. REDMON, S. DIVVALA, R. GIRSHICK et A. FARHADI, « You Only Look Once: Unified, Real-Time Object Detection », en, Las Vegas, NV, USA : IEEE, 2016, p. 779-788.

[113] O. RONNEBERGER, P. FISCHER et T. BROX, « U-Net: Convolutional Networks for Biomedical Image Segmentation », en, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, t. 9351, Munich, Germany : Springer, 2015.

[114] S. A. CHERAGHI, V. NAMBOODIRI et L. WALKER, « GuideBeacon: Beacon-based indoor wayfinding for the blind, visually impaired, and disoriented », en, in *Pervasive Computing and Communications (PerCom)*, Kona, Big Island, HI, USA : IEEE, 2017, p. 121-130.

[115] P. A. ZIENTARA, S. LEE, G. H. SMITH et al., « Third Eye: A Shopping Assistant for the Visually Impaired », en, *Computer*, t. 50, n° 2, p. 16-24, 2017.

[116] X. LI, H. CUI, J.-R. RIZZO, E. WONG et Y. FANG, « Cross-Safe: A Computer Vision-Based Approach to Make All Intersection-Related Pedestrian Signals Accessible for the Visually Impaired », en, in *Advances in Computer Vision*, t. 944, Cham : Springer International Publishing, 2020, p. 132-146.

[117] N. PARIKH, I. SHAH et S. VAHORA, « Android Smartphone Based Visual Object Recognition for Visually Impaired Using Deep Learning », en, in *International Conference on Communication and Signal Processing (ICCSP)*, Chennai : IEEE, 2018, p. 0420-0425.

[118] F. S. BASHIRI, E. LAROSE, J. C. BADGER, R. M. D'SOUZA, Z. YU et P. PEISSIG, « Object Detection to Assist Visually Impaired People: A Deep Neural Network Adventure », en, in *Advances in Visual Computing*, G. BEBIS, R. BOYLE, B. PARVIN et al., éd., t. 11241, Series Title: Lecture Notes in Computer Science, Cham : Springer International Publishing, 2018, p. 500-510.

[119] K. HE, X. ZHANG, S. REN et J. SUN, « Deep Residual Learning for Image Recognition », en, in *Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA : IEEE, 2016, p. 770-778.

[120] M. SANDLER, A. HOWARD, M. ZHU, A. ZHMOGINOV et L.-C. CHEN, « MobileNetV2: Inverted Residuals and Linear Bottlenecks », en, in *Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA : IEEE, 2018, p. 4510-4520.

[121] M. TAN et Q. V. LE, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, en, arXiv:1905.11946 [cs, stat], 2020.

[122] S. KANIMOZHI, G. GAYATHRI et T. MALA, « Multiple Real-time object identification using Single shot Multi-Box detection », en, in *International Conference on Computational Intelligence in Data Science (ICCIDS)*, Chennai, India : IEEE, 2019, p. 1-5.

[123] S. DUMAN, A. ELEWI et Z. YETGIN, « Design and Implementation of an Embedded Real-Time System for Guiding Visually Impaired Individuals », en, in *International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey : IEEE, 2019, p. 1-5.

[124] H. SON et J. WEILAND, « Wearable System to Guide Crosswalk Navigation for People With Visual Impairment », en, *Frontiers in Electronics*, t. 2, p. 790 081, 2022.

[125] C. TON, A. OMAR, V. SZEDENKO et al., « LIDAR Assist Spatial Sensing for the Visually Impaired and Performance Analysis », en, *Neural Systems and Rehabilitation Engineering*, t. 26, n° 9, p. 1727-1734, 2018.

[126] M. AMBARD, « Software Design for Low-Latency Visuo-Auditory Sensory Substitution on Mobile Devices », en, *Computer and Information Science*, t. 10, n° 2, p. 1, 2017.

[127] M. R. U. SAPUTRA, WIDYAWAN et P. I. SANTOSA, « Obstacle Avoidance for Visually Impaired Using Auto-Adaptive Thresholding on Kinect's Depth Image », en, in *Ubiquitous Intelligence and Computing and Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications*, Bali, Indonesia : IEEE, 2014, p. 337-342.

[128] O. YOUNIS, W. AL-NUAIMY, F. ROWE et M. ALOMARI, « A Smart Context-Aware Hazard Attention System to Help People with Peripheral Vision Loss », en, *Sensors*, t. 19, n° 7, p. 1630, 2019.

[129] J. SESSNER, M. SCHMID, M. LAUER-SCHMALZ et J. FRANKE, « Path Segmentation with Artificial Neural Networks in Low Structured Environments for the Navigation of Visually Impaired People », en, in *Biomedical Robotics and Biomechatronics (BioRob)*, New York City, NY, USA : IEEE, 2020, p. 1242-1247.

[130] S. CARAIMAN, O. ZVORISTEANU, A. BURLACU et P. HERGHELEGIU, « Stereo Vision Based Sensory Substitution for the Visually Impaired », en, *Sensors*, t. 19, n° 12, p. 2771, 2019.

[131] S. CARAIMAN, A. MORAR, M. OWCZAREK et al., « Computer Vision for the Visually Impaired: the Sound of Vision System », en, in *International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy : IEEE, 2017, p. 1480-1489.

[132] Y. LIN, K. WANG, W. YI et S. LIAN, « Deep Learning Based Wearable Assistive System for Visually Impaired People », en, in *International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South) : IEEE, 2019, p. 2549-2557.

[133] K. PARK, Y. OH, S. HAM et al., « SideGuide:A Large-scale Sidewalk Dataset for Guiding Impaired People », en, in *International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA : IEEE, 2020, p. 10 022-10 029.

[134] T.-Y. LIN, M. MAIRE, S. BELONGIE et al., « Microsoft COCO: Common Objects in Context », en, in *European Conference on Computer Vision (ECCV)*, arXiv:1405.0312 [cs], t. 8693, Springer, 2014, p. 740-755.

[135] F. YU, H. CHEN, X. WANG et al., « BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning », en, Seattle, WA, USA, 2020, p. 2633-2642.

[136] J. GEYER, Y. KASSAHUN, M. MAHMUDI et al., *A2D2: Audi Autonomous Driving Dataset*, en, arXiv:2004.06320 [cs, eess], 2020.

[137] J. XIAO, J. HAYS, K. A. EHINGER, A. OLIVA et A. TORRALBA, « SUN database: Large-scale scene recognition from abbey to zoo », en, in *Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA : IEEE, 2010, p. 3485-3492.

[138] B. ZHOU, H. ZHAO, X. PUIG et al., « Semantic Understanding of Scenes through the ADE20K Dataset », en, *International Journal of Computer Vision*, t. 127, p. 302-321, 2018, arXiv:1608.05442 [cs].

[139] A. DAI, A. X. CHANG, M. SAVVA, M. HALBER, T. FUNKHOUSER et M. NIESSNER, « ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes », en, in *Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI : IEEE, 2017, p. 2432-2443.

[140] H. CAESAR, J. UIJLINGS et V. FERRARI, « COCO-Stuff: Thing and Stuff Classes in Context », en, in *Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA : IEEE, 2018, p. 1209-1218.

[141] X. HUANG, P. WANG, X. CHENG, D. ZHOU, Q. GENG et R. YANG, « The ApolloScape Open Dataset for Autonomous Driving and its Application », en, *Pattern Analysis and Machine Intelligence*, t. 42, nº 10, p. 2702-2719, 2020, arXiv:1803.06184 [cs].

[142] M. CORDTS, M. OMRAN, S. RAMOS et al., « The Cityscapes Dataset for Semantic Urban Scene Understanding », en, in *Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA : IEEE, 2016, p. 3213-3223.

[143] G. NEUHOLD, T. OLLMANN, S. R. BULO et P. KONTSCHIEDER, « The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes », en, in *International Conference on Computer Vision (ICCV)*, Venice : IEEE, 2017, p. 5000-5009.

[144] M. MASHIATA, T. ALI, P. DAS et al., « Towards assisting visually impaired individuals: A review on current status and future prospects », en, *Biosensors and Bioelectronics: X*, t. 12, p. 100 265, 2022.

[145] S. ZAFAR, M. ASIF, M. B. AHMAD et al., « Assistive Devices Analysis for Visually Impaired Persons: A Review on Taxonomy », en, *IEEE Access*, t. 10, p. 13 354-13 366, 2022.

[146] R. VELÁZQUEZ, E. PISSALOUX, P. RODRIGO, M. CARRASCO, N. GIANNOCCARO et A. LAY-EKUAKILLE, « An Outdoor Navigation System for Blind Pedestrians Using GPS and Tactile-Foot Feedback », en, *Applied Sciences*, t. 8, nº 4, p. 578, 2018, Number: 4.

[147] J. BAI, D. LIU, G. SU et Z. FU, « A Cloud and Vision-based Navigation System Used for Blind People », en, in *Artificial Intelligence, Automation and Control Technologies (AIACT)*, Wuhan, China : ACM Press, 2017, p. 1-6.

[148] V. NAIR, M. BUDHAI, G. OLMSCHENK, W. H. SEIPLE et Z. ZHU, « ASSIST: Personalized Indoor Navigation via Multimodal Sensors and High-Level Semantic Information », en, in *European Conference on Computer Vision (ECCV) Workshops*, L. LEAL-TAIXÉ et S. ROTH, éd., t. 11134, Series Title: Lecture Notes in Computer Science, Cham : Springer International Publishing, 2019, p. 128-143.

[149] V. ALGAZI, R. DUDA, D. THOMPSON et C. AVENDANO, « The CIPIC HRTF database », en, in *Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, USA : IEEE, 2001, p. 99-102.

[150] J. SODNIK, R. SUŠNIK, M. ŠTULAR et S. TOMAŽIČ, « Spatial sound resolution of an interpolated HRIR library », en, *Applied Acoustics*, t. 66, nº 11, p. 1219-1234, 2005.

[151] A. S. BREGMAN, *Auditory scene analysis: hearing in complex environments* (Thinking in Sound: The Cognitive Psychology of Human Audition). Oxford University Press, 1993.

[152] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « Low-Latency Human-Computer Auditory Interface Based on Real-Time Vision Analysis », en, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore : IEEE, 2022, p. 36-40.

[153] C. BORDEAU, F. SCALVINI, C. MIGNIOT, J. DUBOIS et M. AMBARD, « Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution », en, *Frontiers in Psychology*, t. 14, p. 1 079 998, 2023.

[154] M. AMBARD, Y. BENEZETH et P. PFISTER, « Mobile Video-to-Audio Transducer and Motion Detection for Sensory Substitution », en, *Frontiers in ICT*, t. 2, p. 1-13, 2015.

[155] C. TSIRMPAS, A. ROMPAS, O. FOKOU et D. KOUTSOURIS, « An indoor navigation system for visually impaired and elderly people based on Radio Frequency Identification (RFID) », en, *Information Sciences*, t. 320, p. 288-305, 2015.

[156] R. IVANOV, « Indoor navigation system for visually impaired », en, in *Computer Systems and Technologies - CompSysTech*, Sofia, Bulgaria : ACM Press, 2010, p. 143.

[157] M. MARTINEZ, A. ROITBERG, D. KOESTER, R. STIEFELHAGEN et B. SCHAUERTE, « Using Technology Developed for Autonomous Cars to Help Navigate Blind People », en, in *International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy : IEEE, 2017, p. 1424-1432.

[158] B. BENLIGIRAY, C. TOPAL et C. AKINLAR, « STag: A Stable Fiducial Marker System », en, *Image and Vision Computing*, t. 89, p. 158-169, 2019.

[159] E. OLSON, « AprilTag: A robust and flexible visual fiducial system », en, in *International Conference on Robotics and Automation*, Shanghai, China : IEEE, 2011, p. 3400-3407.

[160] F. BERGAMASCO, A. ALBARELLI, E. RODOLA et A. TORSELLO, « RUNE-Tag: A high accuracy fiducial marker with strong occlusion resilience », en, in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA : IEEE, 2011, p. 113-120.

[161] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT, S. ARGON et J. DUBOIS, « Visual-auditory substitution device for indoor navigation based on fast visual marker detection », en, in *Signal-Image Technology & Internet-Based Systems (SITIS)*, Dijon, France : IEEE, 2022, p. 259-266.

[162] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « uB-VisioGeoloc: An Image Sequences Dataset of Pedestrian Navigation Including Geolocalised-Inertial Information and Spatial Sound Rendering of the Urban Environment's Obstacles. », *Data In Brief (accepté)*, p. 1-10, 2023.

[163] N. WOJKE, A. BEWLEY et D. PAULUS, « Simple Online and Realtime Tracking with a Deep Association Metric », en, in *International Conference on Image Processing (ICIP)*, arXiv:1703.07402 [cs], Beijing, China : IEEE, 2017, p. 3645-3649.

[164] H. W. KUHN, « The Hungarian method for the assignment problem », en, *Naval Research Logistics Quarterly*, t. 2, n° 1-2, p. 83-97, 1955, Number: 1-2.

[165] R. E. KALMAN, « A New Approach to Linear Filtering and Prediction Problems », en, *Journal of Basic Engineering*, t. 82, n° 1, p. 35-45, 1960.

[166] C. YU, J. WANG, C. PENG, C. GAO, G. YU et N. SANG, « BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation », en, in *European Conference on Computer Vision (ECCV)*, t. 11217, Munich, Germany : Springer, 2018, p. 325-341.

[167] C. YU, C. GAO, J. WANG, G. YU, C. SHEN et N. SANG, « BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation », en, *International Journal of Computer Vision*, t. 129, p. 3051-3068, 2021.

[168] M. FAN, S. LAI, J. HUANG et al., « Rethinking BiSeNet For Real-time Semantic Segmentation », en, in *Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA : IEEE, 2021, p. 9711-9720.

[169] J. PENG, Y. LIU, S. TANG et al., « PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model », en, p. 1-8, 2022.

[170] Y. HONG, H. PAN, W. SUN et Y. JIA, « Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes », en, p. 1-12, 2021.

[171] E. W. DIJKSTRA, L. BEAUGUITTE et M. MAISONOBE, « E.W. Dijkstra, 1959, A Note on Two Problems in Connexion with Graphs. Numerische Mathematik 1, p. 269271 Version bilingue et commentée », en,

[172] P. HART, N. NILSSON et B. RAPHAEL, « A Formal Basis for the Heuristic Determination of Minimum Cost Paths », *Systems Science and Cybernetics*, t. 4, n° 2, p. 100-107, 1968, Number: 2.

[173] J. DUBOIS et M. MATTAVELLI, « Embedded co-processor architecture for CMOS based image acquisition », en, in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, t. 3, Barcelona, Spain : IEEE, 2003, p. II–591-4.

[174] K. KHATTAB, J. DUBOIS et J. MITERAN, « Cascade Boosting-Based Object Detection from High-Level Description to Hardware Implementation », en, *EURASIP Journal on Embedded Systems*, t. 2009, p. 1-12, 2009.

[175] B. SENOUCI, I. CHARFI, B. HEYRMAN, J. DUBOIS et J. MITERAN, « Fast prototyping of a SoC-based smart-camera: a real-time fall detection case study », en, *Journal of Real-Time Image Processing*, t. 12, n° 4, p. 649-662, 2016.

[176] R. THAVOT, R. MOSQUERON, J. DUBOIS et M. MATTAVELLI, « Generation of Hardware/Software Systems Based on CAL Dataflow Description », en, in *Algorithm-Architecture Matching for Signal and Image Processing*, t. 73, Series Title: Lecture Notes in Electrical Engineering, Dordrecht : Springer Netherlands, 2011, p. 275-292.

[177] K. ABDELOUAHAB, M. PELCAT, J. SEROT, C. BOURRASSET, J.-C. QUINTON et F. BERRY, *Hardware Automated Dataflow Deployment of CNNs*, en, arXiv:1705.04543 [cs], 2017.

[178] M. MARTELLI, « Approche haut niveau pour l'accélération d'algorithmes sur des architectures hétérogènes CPU/GPU/FPGA. Application à la qualification des radars et des systèmes d'écoute électromagnétique », thèse de doct., 2019.

[179] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « Outdoor Navigation Assistive System Based on Robust and Real-Time Visual–Auditory Substitution Approach », en, *Sensors*, t. 24, n° 1, p. 166, 2023.

[180] C. BORDEAU, F. SCALVINI, C. MIGNIOT, J. DUBOIS et M. AMBARD, « Distance perception of objects using visual-to-auditory sensory substitution: comparison of conversion methods based on sound intensity and envelope modulation », in *Auditory Perception, Cognition, & Action Meeting (APCAM)*, Boston, MA, 2022, p. 1.

[181] F. SCALVINI, C. BORDEAU, M. AMBARD, C. MIGNIOT et J. DUBOIS, « Système d'assistance à la mobilité en milieu urbain des personnes malvoyantes via une substitution de l'information visuelle par un signal auditif », in *Compas*, t. Poster, Annecy, France, 2023.