Université de Montréal

Reconnaissance de postures humaines par fusion de la silhouette et de l'ombre dans l'infrarouge

par
Rafik Gouiaa

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Décembre, 2016

RÉSUMÉ

Les systèmes multicaméras utilisés pour la vidéosurveillance sont complexes, lourds et coûteux. Pour la surveillance d'une pièce, serait-il possible de les remplacer par un système beaucoup plus simple utilisant une seule caméra et une ou plusieurs sources lumineuses en misant sur les ombres projetées pour obtenir de l'information 3D ?

Malgré les résultats intéressants offerts par les systèmes multicaméras, la quantité d'information à traiter et leur complexité limitent grandement leur usage. Dans le même contexte, nous proposons de simplifier ces systèmes en remplaçant une caméra par une source lumineuse. En effet, une source lumineuse peut être vue comme une caméra qui génère une image d'ombre révélant l'objet qui bloque la lumière. Notre système sera composé par une seule caméra et une ou plusieurs sources lumineuses infrarouges (invisibles à l'oeil). Malgré les difficultés prévues quant à l'extraction de l'ombre et la déformation et l'occultation de l'ombre par des obstacles (murs, meubles...), les gains sont multiples en utilisant notre système. En effet, on peut éviter ainsi les problèmes de synchronisation et de calibrage de caméras et réduire le coût en remplaçant des caméras par de simples sources infrarouges.

Nous proposons deux approches différentes pour automatiser la reconnaissance de postures humaines. La première approche reconstruit la forme 3D d'une personne pour faire la reconnaissance de la posture en utilisant des descripteurs de forme. La deuxième approche combine directement l'information 2D (ombre+silhouette) pour faire la reconnaissance de postures.

Scientifiquement, nous cherchons à prouver que l'information offerte par une silhouette et l'ombre générée par une source lumineuse est suffisante pour permettre la reconnaissance de postures humaines élémentaires (p.ex. debout, assise, couchée, penchée, etc.).

Le système proposé peut être utilisé pour la vidéosurveillance d'endroits non encombrés tels qu'un corridor dans une résidence de personnes âgées (pour la détection des chutes p. ex.) ou d'une compagnie (pour la sécurité). Son faible coût permettrait un

plus grand usage de la vidéosurveillance au bénéfice de la société. Au niveau scientifique, la démonstration théorique et pratique d'un tel système est originale et offre un grand potentiel pour la vidéosurveillance.

Mots clés: Reconnaissance de postures humaines, capteur infrarouge, calibrage de caméra, transfert d'apprentissage, apprentissage machine, reconstruction 3D, images synthétiques, réseau de neurones convolutionnel.

ABSTRACT

Human posture recognition (HPR) from video sequences is one of the major active research areas of computer vision. It is one step of the global process of human activity recognition (HAR) for behaviors analysis. Many HPR application systems have been developed including video surveillance, human-machine interaction, and the video retrieval. Generally, applications related to HPR can be achieved using mainly two approaches : single camera or multi-cameras. Despite the interesting performance achieved by multi-camera systems, their complexity and the huge information to be processed greatly limit their widespread use for HPR.

The main goal of this thesis is to simplify the multi-camera system by replacing a camera by a light source. In fact, a light source can be seen as a virtual camera, which generates a cast shadow image representing the silhouette of the person that blocks the light. Our system will consist of a single camera and one or more infrared light sources. Despite some technical difficulties in cast shadow segmentation and cast shadow deformation because of walls and furniture, different advantages can be achieved by using our system. Indeed, we can avoid the synchronization and calibration problems of multiple cameras, reducing the cost of the system and the amount of processed data by replacing a camera by one light source.

We introduce two different approaches in order to automatically recognize human postures. The first approach directly combines the person's silhouette and cast shadow information, and uses 2D silhouette descriptor in order to extract discriminative features useful for HPR. The second approach is inspired from the shape from silhouette technique to reconstruct the visual hull of the posture using a set of cast shadow silhouettes, and extract informative features through 3D shape descriptor. Using these approaches, our goal is to prove the utility of the combination of person's silhouette and cast shadow information for recognizing elementary human postures (stand, bend, crouch, fall,...)

The proposed system can be used for video surveillance of uncluttered areas such as a corridor in a senior's residence (for example, for the detection of falls) or in a company

(for security). Its low cost may allow greater use of video surveillance for the benefit of society.

Keywords: Human posture recognition, infrared sensor, camera calibration, transfer learning, machine learning, 3D reconstruction, synthetic images, convolution neural network.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

1D : Une dimensions (scalaire).

2D : Deux dimensions.

3D : Trois dimensions.

CNN : Convolution Neural network.

DT : Distance Transform.

HAR : Human activity recognition.

HPR : Human posture recognition.

IR : InfraRed.

KNN : K-Nearest Neighbours.

MLP : Multi-Layer Perceptron.

ReLU : Rectified Non Linear function.

SFS : Shpae From Silhouettes.

SVM : Support Vector Machines.

SVD : Singular Values Decomposition.

VH : Visual Hull.

WKNN : Weighted K-Nearest Neighbours.

REMERCIEMENTS

Tout d'abord, je remercie très chaleureusement mon directeur de thèse M. Jean Meunier qui m'a donné la chance de réaliser ce travail dans les meilleures conditions au laboratoire de traitement d'images du Département d'Informatique et de Recherche opérationnelle (DIRO). Je vous remercie d'avoir cru à mes compétences, pour votre excellent encadrement, pour le temps qui m'a accordé malgré vos nombreuses charges et pour vos précieux conseils. Pour tout ce qui vous m'avez fait, je vous remercie très sincèrement.

Je tiens à remercie également le professeur Sébastien Roy pour ses multiples conseils, pour sa disponibilité et d'avoir accepté de rapporter cette thèse.

Je souhaiterais remercier sincèrement les membres du jury pour le temps qu'ils ont accordé à l'examination de ce manuscrit et à l'élaboration de leur rapport.

Un grand merci à toutes les personnes qui ont accepté de m'aider pendant la réalisation de la base de données de postures humaines. Merci pour votre disponibilité et pour votre patience jusqu'à finir l'acquisition.

Merci à tous les membres du laboratoire Image ainsi que les doctorants et les stagiaires que j'ai rencontrés pendant ces années pour leurs comportements et leurs organisations.

Merci à la Mission universitaire de Tunisie en Amérique du Nord (MUTAN) et au gouvernement du Canada pour leur support financier.

Merci à tous les professeurs (Jean Meunier, Michel Boyer, Pascal Vincent, Stefan Monnier,...) de m'avoir confié les démonstrations de plusieurs cours. Cette expérience comme étant auxiliaire d'enseignement m'a permise d'acquérir beaucoup de compétences en matière d'enseignement.

Mes plus vifs remerciements s'adressent à tous les membres de ma famille qui tiennent une place immense dans mon coeur. Je vous remercie très chaleureusement pour votre amour et votre soutien constant.

CHAPITRE 1

INTRODUCTION GÉNÉRALE

1.1 Contexte

La vision par ordinateur est une thématique active et passionnante de recherche dont l'objectif est de développer des algorithmes et des représentations qui permettront à l'ordinateur d'analyser et d'interpréter des informations visuelles de façon autonome. L'une des applications pertinentes dans le contexte de la vision par ordinateur est la reconnaissance de postures humaines. Ceci pourrait être une étape fondamentale du processus global de l'analyse du comportement humain. L'analyse et l'interprétation du mouvement chez l'être humain peuvent être impliqués dans des nombreuses applications notamment la vidéo surveillance intelligente, l'interaction personne-machine et la réadaptation à domicile.

Dans le contexte de la vision par ordinateur, la reconnaissance de la posture humaine et l'analyse du mouvement humain peuvent s'étudier généralement par deux systèmes : le système *monocaméra* et le système *multicaméras*.

- **Système mono-caméra** : Un tel système est composé d'une caméra fixe ou mobile, calibrée ou non, installée dans un environnement interne ou externe afin de capturer les mouvements d'une personne dans une scène, à partir d'un seul angle de vue. Différentes méthodes utilisant une seule caméra ont été proposées pour la reconnaissance de postures humaines, voir par exemple, (Kellokumpu et Heikkilä, 2005) , (Schindler et Van Gool, 2008), Wang et al. (2011) et Ji et al. (2013). Bien que ces méthodes puissent être adaptées dans différents environnements et aient une complexité de temps et de mémoire raisonnable, leurs performances sont dépendantes de point de vue de la capture de postures, et très sensibles aux occultations et ambiguïtés. En outre, les postures 2D récupérées par une caméra perspective ne sont pas souvent précises car plusieurs silhouettes de classes différentes peuvent être très similaires.

- **Système multicaméras** : Formé d'un réseau de caméras souvent calibrées, approximativement synchronisées et installées d'une manière où une personne dans une scène est capturée de différents angles. Dans ce contexte, plusieurs recherches ont été faites pour la reconnaissance de postures humaines, par exemple (Nadia et al., 2008), (Onishi et al., 2008), (Cohen et Li, 2003), (Pellegrini et Iocchi, 2007), (Wu et Aghajan, 2007b). Généralement, en utilisant un ensemble de caméras, la quantité de données visuelles fournies de plusieurs angles de vues permet une meilleure interprétation de la posture. Ainsi, une plus grande précision dans la reconnaissance de postures et une meilleure robustesse à l'occultation sont également obtenues lorsque plusieurs caméras sont utilisées.

1.2 Problématiques posées

Malgré les solutions offertes par les systèmes multicaméras à propos de l'ambiguïté et l'occultation, ainsi que les bons résultats obtenus, leur utilisation reste limitée par rapport aux systèmes monoculaires à cause de différentes difficultés techniques. Parmi ces difficultés, la synchronisation et le calibrage d'un réseau de caméra dont l'objectif principal est de calculer les paramètres géométriques nécessaires pour récupérer l'information de la profondeur perdue lors de la projection perspective. En outre, l'installation d'un ensemble de caméras dans une scène peut être lourd et coûteux. Par exemple, la détection de la chute chez les personnes âgées avec un tel système nécessite l'installation d'un ensemble de caméras dans chaque pièce ce qui n'est pas évident en pratique. Un réseau de caméra produit une énorme quantité de données qui doivent être respectivement affichées, transférées et traitées. Ceci demande l'installation d'un réseau de communication efficace capable d'assurer toutes ces fonctionnalités. ce qui implique d'autres enjeux comme la compression de données.

1.3 Contributions

Pour faire face à ces problèmes et bénéficier des avantages offerts par les deux systèmes, l'idée est de concevoir un système multivues en utilisant une seule caméra et un ensemble de sources lumineuses infrarouges installées autour de la scène. En fait, une source lumineuse peut être vue comme une caméra virtuelle permettant d'offrir une image de l'objet (une personne) bloquant la lumière à travers l'ombre projetée. Par conséquent, les ombres projetées par les différentes sources lumineuses peuvent être considérées comme des images, de la personne, capturées de différents angles de vues. Alors, un tel système nous permet de filmer simultanément la personne à partir de différentes points de vues par le biais d'une seule caméra.

Dans ce contexte, nous proposons deux approches pour automatiser la reconnaissance de postures humaines. La première approche combine directement la silhouette de la personne avec sa silhouette d'ombre pour assurer la reconnaissance de postures. Tandis que, la deuxième approche exploite les différentes silhouettes (ombres+ silhouette de la personne) pour reconstruire la forme 3D du corps sous forme d'une enveloppe convexe (*Visual Hull*) qui est décrit par des descripteurs de forme pour la reconnaissance de postures.

1.4 Structure de thèse

Cette thèse par articles est composé de 6 chapitres incluant deux articles de conférences et un article de journal.

Chapitre 2 : présente les différents outils mathématiques et méthodes informatiques nécessaires pour la reconnaissance de postures humaines qui seront utilisés dans les chapitres subséquents.

Chapitre 3 : présente notre première contribution avec un article de conférence portant sur la reconnaissance de postures humaines en utilisant la combinaison de la silhouette de la personne et de l'ombre invisible projetée par 4 sources lumineuses infrarouges.

Chapitre 4 : étudie notre deuxième contribution portant sur la classification de postures humaines en misant sur l'apprentissage de l'apparence de l'ombre en utilisant un réseau de neuronnes convolutionnel. Cependant, dans ce travail, nous avons considéré une scène plus compliquée dont l'ombre se projette sur le plancher et le mur d'un corridor, en générant des images difficiles à reconnaitre.

Chapitre 5 : expose la troisième contribution avec un article de conférence portant sur la classification de postures humaines en se basant sur la reconstruction 3D du corps humain à partir d'une seule caméra et 4 sources lumineuses infrarouges.

Chapitre 6 : Une conclusion générale résumant nos principales contributions ainsi que diverses perspectives de recherche.

1.5 Publications

- R. Gouiaa, J.Meunier, *Learning cast shadow appearance for human posture recognition* (submitted to Pattern Recognition Letters Nov 2016.)
- R. Gouiaa, J.Meunier, *Human Posture Recognition By combining Silhouette and Infrared Cast shadows*, IPTA 2015.
- R.Gouiaa, J. Meunier, *Human Posture Classification based on 3D body Shape recovered using Silhouette and Cast shadows*, IPTA 2015.
- R. Gouiaa, J.Meunier, *3D reconstruction by fusioning shadow and silhouette information*. Computer and Robot Vision, 2014.

CHAPITRE 2

OUTILS MATHÉMATIQUES ET MÉTHODES INFORMATIQUES POUR LA RECONNAISSANCE DE LA POSTURE HUMAINE

En vision par ordinateur, une caméra perspective est représentée généralement par le modèle sténopé (ou le modèle pin-hole dans la littérature anglo-saxonne) avec ou sans distorsion. Il s'agit d'une modélisation simple et linéaire du processus de formation des images au sein d'une caméra. C'est le modèle le plus couramment utilisé dans le domaine de vision par ordinateur.

Le calibrage d'une caméra consiste à estimer les paramètres du modèle mathématique (sténopé) qui la représente. Selon ce modèle, une caméra possède deux types de paramètres : les paramètres intrinsèques (la distance focale, les facteurs d'agrandissement de l'image, les coordonnées de centre optique de la caméra...) qui sont internes à la caméra, et les paramètres extrinsèques qui décrivent sa position et son orientation dans l'espace 3D.

2.1 La modélisation géométrique d'une caméra et la formation d'image

Une caméra perspective, selon le modèle sténopé, sert à transformer un point 3D $M = (X, Y, Z)$ de l'espace en un point d'image $m = (u, v)$ en exerçant trois transformations élémentaires comme illustrées sur la figure 2.1 : ① la transformation entre le repère monde et le repère caméra, ② la transformation entre le repère caméra et le repère capteur, ③ la transformation entre le repère capteur et le repère image.

1. ***Transformation entre le repère monde et le repère caméra :***
   La transformation ① permet de passer du repère 3D R_w au repère caméra R_c. C'est une transformation rigide représentée par les paramètres extrinsèques de la caméra, c'est à dire sa position et son orientation par rapport au repère de référence. Cette transformation peut s'écrire sous la forme d'une matrice homogène

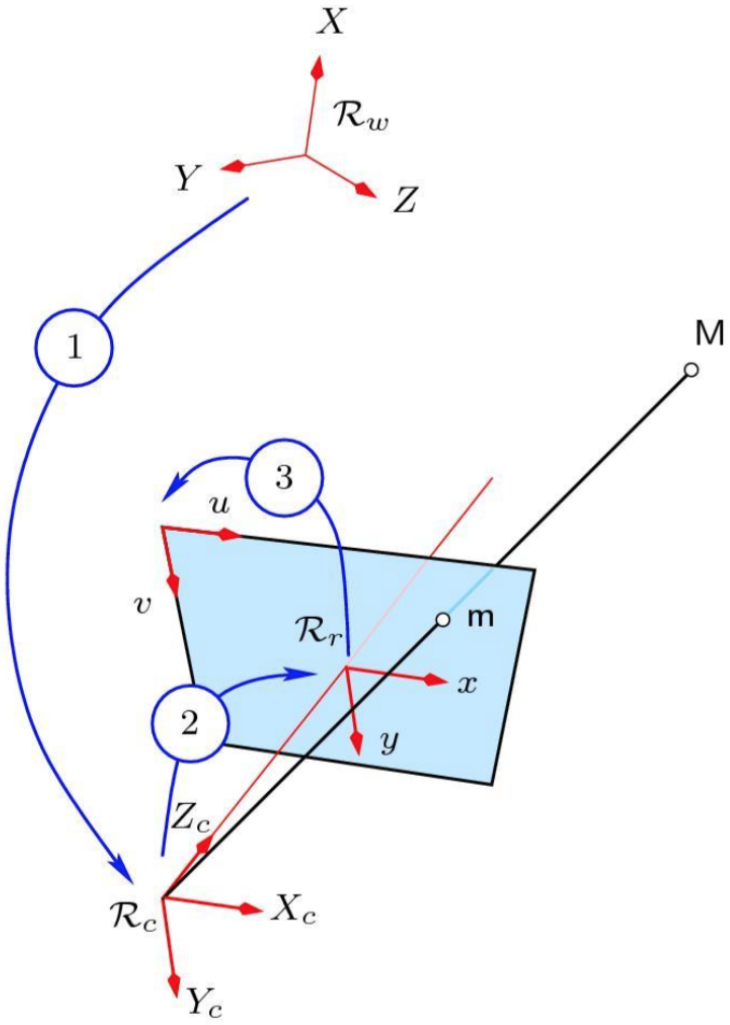


Figure 2.1 – Les trois transformations élementaires et le modèle sténopé.

associant un point de coordonnées homogènes $(X,Y,Z,1)$ dans le repère monde à un point de coordonnées $(X_c,Y_c,Z_c,1)$ dans le repère caméra R_c (voir l'équation 2.1).

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & T_x \\ r_{21} & r_{22} & r_{23} & T_y \\ r_{31} & r_{32} & r_{33} & T_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = M_{ext} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.1}$$

où $r_{i,j}$ représentent les composantes de la matrice de rotation de la caméra,

T_x, T_y, T_z représentent la translation de la caméra et M_{ext} est la matrice externe de la caméra.

2. ***Transformation entre le repère caméra et le repère capteur :***
   C'est une transformation perspective (une matrice de 3×4 notée P) permettant de projeter un point 3D $(X_c, Y_c, Z_c, 1)$ du repère caméra à un point 2D $(x, y, 1) = (f \cdot \frac{X_c}{Z_c}, f \cdot \frac{X_c}{Z_c}, 1)$ dans le plan image (en unité métrique). Tel qu ?illustré par ② sur la figure 2.1, cette transformation peut s'écrire en coordonnées homogènes comme suit :

$$s \times \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = P \cdot \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{2.2}$$

   où f est le focal et s est un facteur d'échelle (géométrie projective).

3. ***Transformation entre le repère capteur et le repère image***
   C'est la dernière étape dans le processus de formation d'image, notée ③ sur la figure 2.1. Elle explique le passage d'un point 2D (x, y) de coordonnées en unité métrique (continue) en un point de coordonnées (u, v) en unité pixel (discrète). Cette conversion est faite via les paramètres intrinsèques de la caméra.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = M_{int} \cdot \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{2.3}$$

   où M_{int} est la matrice interne de la caméra, $\alpha_x = f \cdot m_x, \alpha_y = f \cdot m_y$, et m_x et m_y sont le nombre de pixels par unité de longueur métrique, respectivement, dans la direction x et y du capteur CCD. (u_0, v_0) est l'origine de coordonnées du plan image.

D'après les équations 2.1 et 2.3, la matrice de la caméra perspective s'écrit sous la forme suivante :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M_{int} \cdot M_{ext} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.4}$$

2.2 Correction de la distorsion

Le système de formation d'image basé sur le modèle sténopé et décrit par l'équation 2.4 représente une caméra idéale, c'est-à-dire que la caméra obéit parfaitement à une transformation linéaire de l'image, par exemple, une ligne droite 3D reste droite après sa projection sur le plan image 2D. Cependant, en réalité, les systèmes optiques ne sont pas linéaires et possèdent de la distorsion qui se traduit par la courbure des lignes droites de l'objet filmé.

La distorsion se manifeste dans la phase ② de la formation d'image où l'équation 2.2 n'est plus linéaire et les coordonnées idéales (x, y) sont reliées aux coordonnées réelles $(\breve{x}, \breve{y})$ par l'équation suivante :

$$\begin{cases} \breve{x} = x + \delta_{r_x} + \delta_{t_x} \\ \breve{y} = y + \delta_{r_y} + \delta_{t_y} \end{cases} \tag{2.5}$$

où $(\delta_{r_x}, \delta_{t_x})$ et $(\delta_{r_y}, \delta_{t_y})$ sont les deux composantes de distorsions radiales et tangentielles respectivement suivant les axes x et y.

1. **la distorsion radiale** : Elle est symétrique par rapport à l'axe optique, et principalement causé par la courbure des lentilles et la position du diaphragme vis-à-vis du centre optique. Cette distorsion provoque un déplacement à l'extérieur ou à l'intérieur d'un point d'image par rapport à sa position idéale. Ceci donne respectivement une image en forme de barillet ou coussinet (voir figure 2.2). D'après (Heikkila et Silvén, 1997), La distorsion radiale peut être approximée par l'ex-

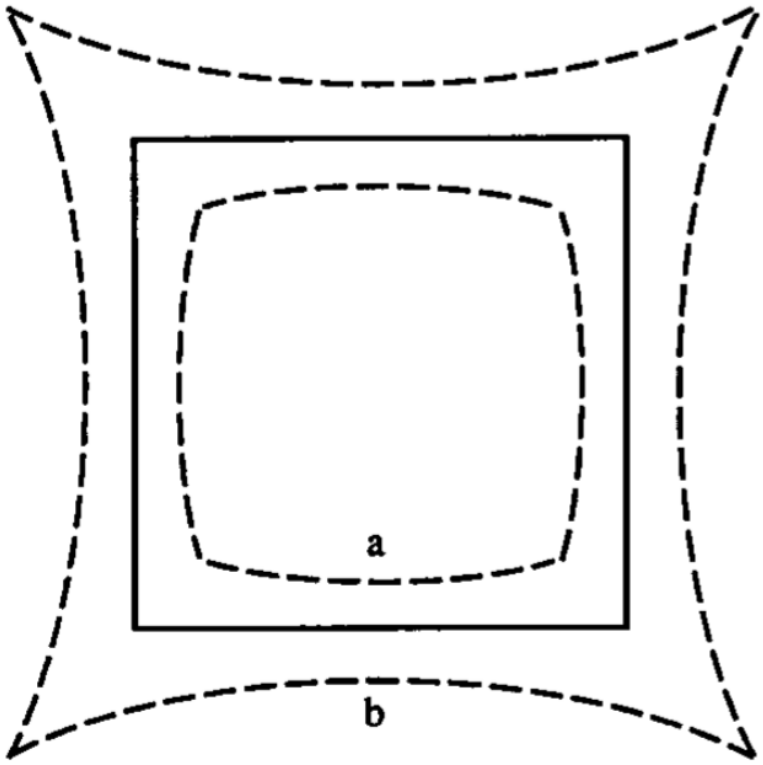


Figure 2.2 – Effet de la distorsion radiale : a) distorsion en barillet, b) distorsion en coussinet (Weng et al., 1992)

pression suivante :

$$\begin{cases} \delta_{r_x} = x(1 + k_1 r^2 + k_2 r^4 + \dots) \\ \delta_{r_y} = y(1 + k_1 r^2 + k_2 r^4 + \dots) \end{cases} \tag{2.6}$$

où k_1 et k_2 sont les coefficients de la distorsion et $r = \sqrt{x^2 + y^2}$.

2. **la distorsion tangentielle :** elle est due à deux défauts majeurs : Il se peut que la lentille ne soit pas positionnée perpendiculairement à l'axe principal ou bien que la lentille ne soit pas centrée par rapport à l'axe principal. D'après (Heikkila et Silvén, 1997), la distorsion tangentielle est modélisée par l'équation suivante :

$$\begin{cases} \delta_{t_x} = 2p_1 xy + p_2(r^2 + 2x^2) \\ \delta_{t_y} = p_1(r^2 + 2y^2) + 2p_2 xy \end{cases} \tag{2.7}$$

p_1 et p_2 sont les coefficients de la distorsion tangentielle.

Généralement, l'effet de la distorsion tangentielle est négligeable par rapport à la distorsion radiale et par conséquent la plupart des techniques de calibrage ne la considèrent pas.

2.3 Quelques techniques de calibrage d'une caméra

Le calibrage d'une caméra a été traité intensivement dans la littérature. Dans ce manuscrit, nous citons seulement les deux techniques les plus utilisées :

1. **Calibrage avec un objet 3D :** dans cette approche, généralement, l'objet de calibrage prend la forme d'une mire constituée par deux ou trois plans, deux à deux orthogonaux. L'estimation de paramètres de la caméra est effectuée en observant un ensemble de points précis (coins, marqueurs...) sur l'objet de calibrage. Les coordonnées 3D de ces points sont connues, alors que leurs coordonnées 2D sont mesurées. Une mise en correspondance entre les deux ensembles des coordonnées nous permet d'identifier les paramètres de la caméra, en résolvant un système linéaire surdéterminé (Weng et al., 1992) obtenu à partir de l'équation 2.4.
2. **Calibrage avec un objet 2D :** un objet de calibrage 2D est observé plusieurs fois (au moins deux fois) par la caméra sous différentes orientations. À partir de chaque vue, une homographie reliant les points 3D avec ceux en 2D est calculée. Celle-ci contient de l'information sur les paramètres de la caméra. Dans ce cadre, (Zhang, 2000) a proposé une méthode très populaire dans le milieu de la vision par ordinateur en utilisant un damier comme objet de calibrage.

2.4 Localisation d'une source lumineuse

L'estimation de l'emplacement d'une source lumineuse est un problème important pour diverses applications de vision et d'infographie, telles que la reconstruction d'une forme à partir de l'ombrage (*shape from shading*), la reconnaissance d'objet, le rendu

basé sur l'image et la réalité augmentée. Dans notre cas, nous avions besoin de localiser la source lumineuse afin de récupérer la forme 3D du corps humain (voir chapitre 5).

Dans la littérature, (Bunteong et Chotikakamthorn, 2016) ont utilisé des points caractéristiques dans des ombres projetées (*cast shadows*) pour estimer les positions de sources lumineuses à partir des directions de source estimée en utilisant des reflets spéculaires. Dans (Powell et al., 2000), deux ou trois sphères sont utilisées, et les propriétés spéculaires d'une surface brillante sont utilisées pour trianguler les emplacements des sources lumineuses.

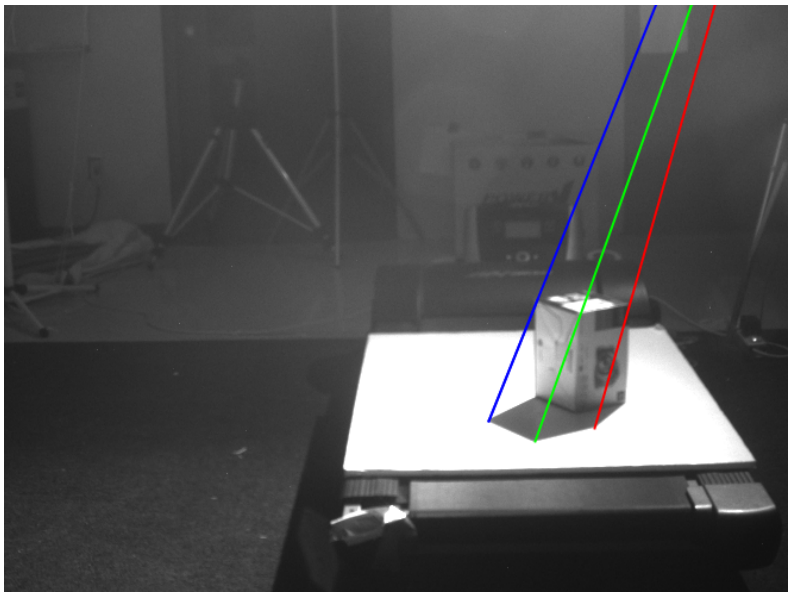


Figure 2.3 – Localisation d'une source lumineuse

Dans notre cas, nous exploitons l'ombre projetée (cast shadow) d'un objet 3D pour estimer la position de la source lumineuse (Gouiaa et Meunier, 2014a). Étant donné un objet 3D tel qu'une boîte en carton, nous définissons un rayon comme la ligne 3D joignant un point 3D (X,Y,Z) (un coin de la boîte) et son point correspondant dans l'ombre projetée (X_s,Y_s,Z_s) (voir la figure 2.3). L'équation d'un rayon peut s'écrire comme suit :

$$r(t) = (X,Y,Z) + ((X_s,Y_s,Z_s) - (X,Y,Z)) \times t. \tag{2.8}$$

Le centre d'une source lumineuse ponctuelle peut être déterminé comme l'intersection

entre au moins deux rayons. Cependant, en pratique, la source lumineuse n'est pas parfaitement ponctuelle. Par conséquent, on ne peut obtenir qu'une approximation d'une position de source lumineuse. Étant donnés, n rayons $(r_1(t1), r_2(t2), \ldots, r_n(tn))$ et un point arbitraire (X_c, Y_c, Z_c), si nous minimisons la somme des carrés des distances d'un point (X_c, Y_c, Z_c) à un point sur chacun des rayons, on trouvera un point qui est, dans le sens de moindres carrés, le plus proche du centre de la source lumineuse. Ainsi, le centre de source de lumière noté S est approximé par l'équation suivante :

$$S = \underset{\{X_c, Y_c, Z_c, t_1, \ldots, t_n\}}{\operatorname{argmin}} \sum_{i=1}^{n} (r_i(t_i) - (X_c, Y_c, Z_c))^2 \tag{2.9}$$

2.5 Descripteur de formes

D'une manière générale, un descripteur de forme est une représentation d'une forme 2D ou 3D par un vecteur contenant un ensemble de valeurs numériques ou une structure graphique servant à décrire la forme géométriquement ou topologiquement (Akgül, 2007). Les descripteurs de forme sont évalués sur la base de plusieurs caractéristiques qui définissent la qualité globale et l'efficacité d'un descripteur de forme. Un descripteur efficace de forme doit avoir certaines caractéristiques communes telles que :

- Discrimination : Distinguer avec précision une forme d'une autre selon des différences subtiles.
- Invariance aux transformations (translation, échelle et rotation).
- Robustesse contre les dégénérescences et la rugosité (erreur d'extraction du fond...).
- Unicité : chaque descripteur de forme doit être uniquement couplé à une forme unique.
- Correspondance partielle : robuste contre les formes incomplètes.
- Insensible au bruit : petits changements de forme ne conduisent pas à de grands changements dans le descripteur de forme.

Les descripteurs de forme peuvent se regrouper en deux catégories : descripteurs 2D et descripteurs 3D (Kazmi et al., 2013).

2.5.1 Descripteurs 2D

Dans la littérature, les descripteurs 2D sont basés essentiellement sur deux approches : contour ou région. Nous présentons ici une revue non exhaustive de ces descripteurs.

Approche contour

Ce sont des descripteurs qui extraient des caractéristiques (*features*) à partir du contour seulement.

Le codage de Freeman est parmi les plus anciennes méthodes utilisées pour la description de contours (Freeman, 1961). Elle consiste à coder les directions du contour dans un repère absolu à partir d'une origine donnée.

(Kopf et al., 2005) ont utilisé le descripteur de courbure CSS (Curvature Scale Space) pour la reconnaissance de postures humaines. CSS divise la forme en segments convexes et concaves en identifiant un ensemble de points d'inflexion ; points où la courbure de la forme est nulle. L'algorithme CSS consiste à calculer la courbure du contour tout en lissant progressivement la courbe puis à finalement générer l'image CSS. (Kellokumpu et Heikkilä, 2005) ont utilisé le descripteur de Fourier pour la reconnaissance des activités humaines à partir d'une séquence des postures. Le descripteur de Fourier est une représentation de la forme obtenue après l'application de la transformée de Fourier sur un contour échantillonné de la forme.

(Mori et Malik, 2002) ont utilisé le contexte de forme (*shape context*) pour estimer la posture de l'être humain. L'idée principale du contexte de forme est de trouver la correspondance ainsi qu'une mesure de dissimilarité entre les deux formes. Pour trouver la correspondance entre les deux formes, N points sont échantillonnés à partir du contour de la forme et un point de référence est fixé. Les points sont échantillonnés en utilisant un algorithme de détection de contour. Ensuite, un ensemble de vecteurs est calculé en partant du point de référence vers tous les autres points échantillonnés. Le contexte de forme pour chaque point p_i est défini comme un histogramme h_i des coordonnées

polaires relatives des points échantillonnés restants.

$$h_i(k) = \#\{Q \neq p_i : (Q - p_i) \in bin(k)\} \tag{2.10}$$

Approche région

Ce sont des descripteurs qui obtiennent des caractéristiques (*features*) à partir de la région entière de la forme.

(Grundmann et al., 2008) ont utilisé la version 3D du descripteur du contexte de forme en combinaison avec la carte de distances (distance transform) pour la reconnaissance des activités humaines. La carte de distance est une image numérique qui associe à chaque pixel de l'image la distance au point du contour (dans une image binaire) le plus proche.

Dans (Sun et al., 2009), les auteurs ont appliqué les moments de Zernike pour extraire deux types de caractéristiques holistiques, l'une est basée sur des images uniques (postures) et l'autre est basée sur l'image d'énergie de mouvement. Enfin, ces caractéristiques ont été utilisées pour la reconnaissance des activités humaines.

(Karahoca et al., 2008) ont extrait des vecteurs numériques de dimension 7 en appliquant les 7 moments de Hu (Hu, 1962) sur des images de mouvement (*Motion history image (MHI)*), pour décrire des activités humaines. Ces vecteurs ont été classifiés par un SVM afin d'inférer l'activité correspondante.

2.5.2 Descripteurs 3D

Les descripteurs de forme 3D ont été largement utilisés pour la reconnaissance de postures humaines. Dans ce manuscrit, nous traitons seulement les descripteurs utilisés pour extraire des caractéristiques à partir d'une forme 3D (le corps humain dans notre cas), tout en étudiant deux grandes techniques utilisées pour la reconstruction d'une forme 3D.

Stéréovision

Brièvement, nous appelons stéréovision, le processus de la reconstruction d'une forme 3D à partir de deux ou plusieurs images d'une même scène, prises sous différents angles de vue. Ce processus consiste en deux phases principales : *la mise en correspondance et la triangulation*

La mise en correspondance est la phase la plus délicate, et aussi celle qui varie le plus d'une méthode à l'autre. Celle-ci consiste à trouver chaque paire de points 2D dans les deux images correspondant au même point 3D dans la scène. Une fois la mise en correspondance est faite pour tous couples de points, ces points peuvent être triangulés afin de générer une carte de disparité (dense ou éparse) représentant les profondeurs des points 3D.

Par exemple, (Sanchez-Riera et al., 2012) ont proposé une méthode de reconnaissance d'activités humaines qui fonctionne avec des vidéos binoculaires. Cette méthode utilise l'approche standard du sac de mots (*bag of words*) sauf que les histogrammes de caractéristiques (histogram of features HOF) ont été construits en utilisant des cartes de disparité.

Dans une autre application, (Pellegrini et Iocchi, 2007) ont présenté un système de suivi et de classification de postures basé sur la vision stéréoscopique. La méthode proposée est basée sur l'appariement des cartes de disparités avec des modèles 3D du corps humain.

Shape from silhouettes

Cette technique, comme son nom indique, utilise des primitives simples, les silhouettes, pour approximer la forme de l'objet (le corps humain dans notre cas). Le *Shape From Silhouettes (SFS)* a été proposé pour la première fois par Baumgart en 1974. Dans sa thèse (Baumgart, 1974), Baumgart a estimé la forme d'une poupée et un cheval jouet en utilisant 4 silhouettes (contours) représentées chacune par un polygone. La forme obtenue est représentée sous la forme d'un polyèdre. Depuis, différentes variantes de *Shape*

From Silhouettes ont été proposées pour la reconstruction d ?une forme 3D. Par exemple, (Kim et Aggarwal, 1986) ont proposé des descriptions volumétriques ("segment de volume" et "parallélépipède rectangle") représentant l'objet estimé à partir d'une séquence de silhouettes contours. Dans (Srinivasan et al., 1990), les auteurs ont développé une technique pour estimer la forme d'objet en intersectant des cônes 3D des silhouettes. (Bacelar et al., 1994) ont développé une méthode pour générer un octree d'un objet à partir d'un ensemble des silhouettes. Pour chaque silhouette projetée sur le plan image, un volume octree 3D est calculé dont chacun de ses octants est déduit par l'intersection avec la silhouette. Le modèle 3D est itérativement reconstruit en intersectant les volumes octree 3D de toutes les silhouettes.

En 1991, (Laurentini, 1994) a proposé un nouveau terme, appelé enveloppe visuelle *visual hull (VH)* . Ce terme est défini comme l'intersection des cônes, formés par les silhouettes et les centres des caméras correspondants. Grâce aux progrès connus au niveau de l'extraction de silhouette, la SFS est devenue une méthode populaire et standard pour l'estimation de la forme humaine et la reconnaissance de postures et l'analyse du mouvement humain. (Peng et Qian, 2011) ont présenté une méthode d'analyse multilinéaire pour extraire des caractéristiques invariantes à la vue de caméra à partir du VH. Ensuite, ces caractéristiques ont été présentées à l'entrée d'un modèle de Markov caché pour déduire la posture. (Auvinet et al., 2011) ont développé une méthode pour la détection de chute chez les personnes âgées. Premièrement, la forme 3D du corps de la personne est construite en utilisant la SFS. Ensuite, une analyse de la distribution verticale du volume est utilisée pour signaler les chutes (voir la figure 2.4).

(Cohen et Li, 2003) ont proposé une approche pour inférer une posture humaine en utilisant un descripteur basé sur l'apparence du corps humain et invariant aux changements des vues. Ce descripteur est obtenu à partir de l'enveloppe visuelle 3D construite à partir d'un ensemble de silhouettes. (Pierobon et al., 2005) ont proposé une méthode de regroupement (*clustering*) basée sur une représentation 3D du corps humain en termes d'enveloppe visuelle. Les caractéristiques ont été représentées par un descripteur de forme calculé image par image et adaptées pour être indépendantes de la position, de

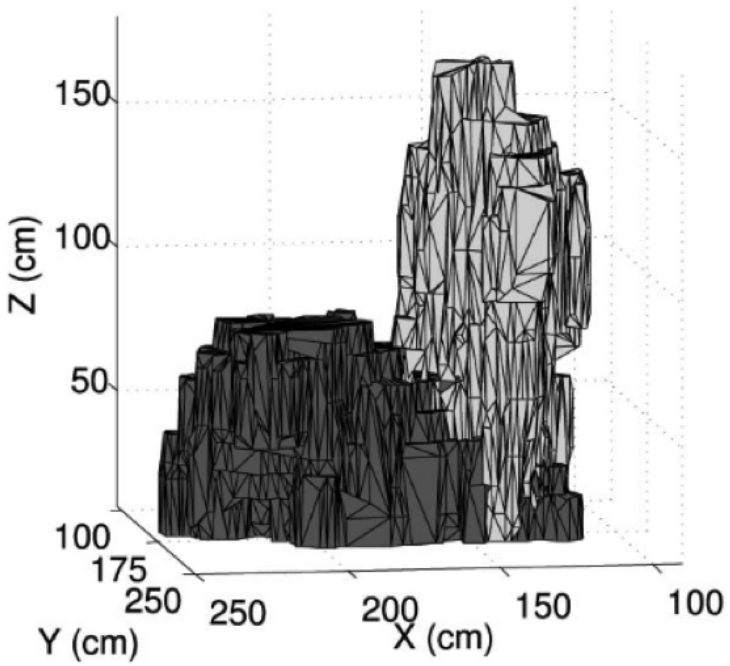


Figure 2.4 – Distribution de volume d'une reconstruction 3D de deux postures : posture debout (gris clair) et posture chute (gris foncé)

la taille, de l'échelle, des proportions corporelles et, éventuellement, être invariant aux rotations.

Principe général de Shape From Silhouettes : étant donné un objet 3D O filmé par K caméras. Soit $\{S_j^k\}$ avec k=1,...,K, les images silhouettes générées par la caméra k à l'instant t_j. Supposons que les caméras sont calibrées, et P^k et C^k sont respectivement la matrice de la projection et le centre optique de la k^{ème} caméra. Ainsi, les coordonnées 2D du point 3D X dans la k^{ème} image est noté comme $x = P^k X$. Par extension de cette dernière notation, $P^k A$ est l'image d'un volume A dans le plan image de la k^{ème} caméra. Inversement, pour chaque caméra et chaque silhouette, un volume A contenant l'objet 3D est défini. Ce volume est appelé un cône visuel. Une illustration d'un scénario de *Shape From Silhouette* est bien expliquée dans la figure 2.5. Soient les définitions suivantes (Cheung, 2003) :

- Un point 3D X de l'objet O est dit consistent avec la silhouette S_j^k donnée par la

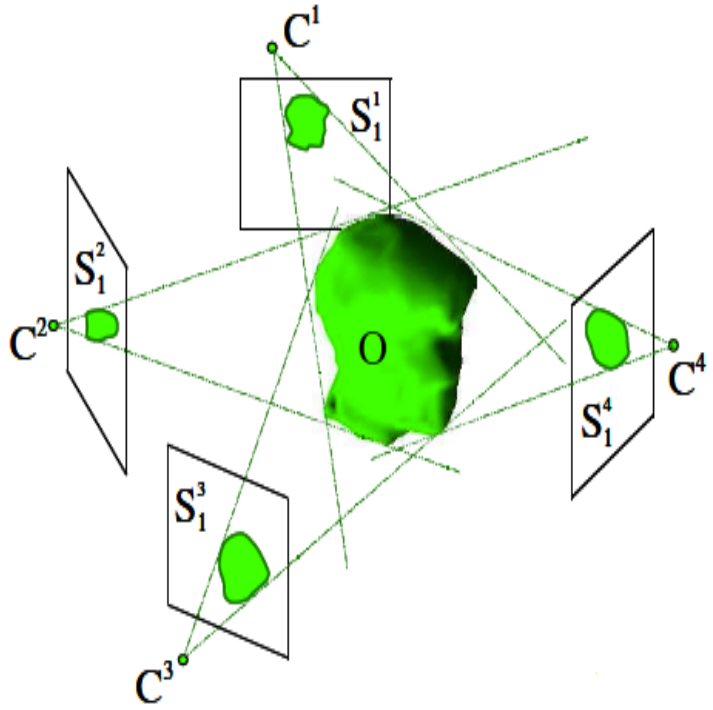


Figure 2.5 – Un exemple d'un scénario de *Shape From Silhouettes*. Un objet O est entouré par 4 caméras. S_1^k et C^k représentent respectivement les silhouettes et le centre de la k^{ème} caméra (Cheung, 2003).

k^{ème} caméra à l'instant j, si et seulement si $P^k X$ appartient à S_j^k.

- Un point 3D X est dit consistent avec les silhouettes si et seulement si pour toutes caméras X est consistent avec S_j^k $(k = 1, \ldots, K)$.
- Une silhouette est dite consistante s'il existe au moins un volume A non vide qui représente la silhouette sinon, on dit qu'elle est inconsistante.

Étant donné K silhouettes $S_j^k, k = 1, \ldots, K$ consistantes, le *visual hull* H_j est défini comme l'intersection des K cônes visuels (Cheung, 2003). Autrement dit, le *visual hull* H_j est le plus grand volume qui représente exactement les K silhouettes consistantes. On dit que H_j est exactement égal à l'objet 3D si pour chaque caméra k, sa projection dans le plan image coïncide avec la silhouette S_j^k. Cependant, celle-ci n'est pas valide en pratique, car les silhouettes ne sont pas consistantes à cause du bruit et des erreurs de calibrage.

Approches de construction d'un visual hull

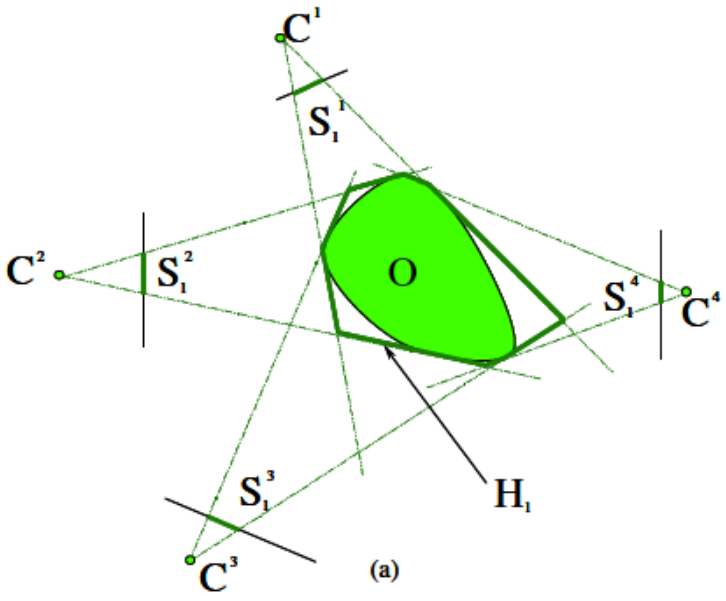


Figure 2.6 – Un exemple 2D de la construction d'un *visual hull* par l'approche surfacique (Cheung, 2003)

- **Approche surfacique :** le *visual hull* est obtenu, d'une manière explicite, par l'intersection des cônes visuels formés par les silhouettes et les centres des caméras. Parmi les travaux qui se basent sur les approches surfaciques, Baumgart (Baumgart, 1974) a été parmi les premiers qui a estimé la forme de l'objet par un polyèdre. Ce dernier est calculé explicitement à partir de l'intersection des formes polygonales décrivant les contours des silhouettes. Un exemple de l'estimation d'un objet en utilisant le *Shape From Silhouettes* surfacique est illustré par la figure 2.6.
- **Approche volumique :** L'espace 3D d'intérêt est divisé en une grille discrète de voxels. Ensuite, l'appartenance de chaque voxel au *visual hull* est vérifiée. Ainsi, un voxel est identifié comme appartenant au volume, si sa projection dans tous les K plans images fait partie de la silhouette correspondante. Enfin, les voxels validés sont dits silhouettes consistantes et forment une estimation du *visual hull*.

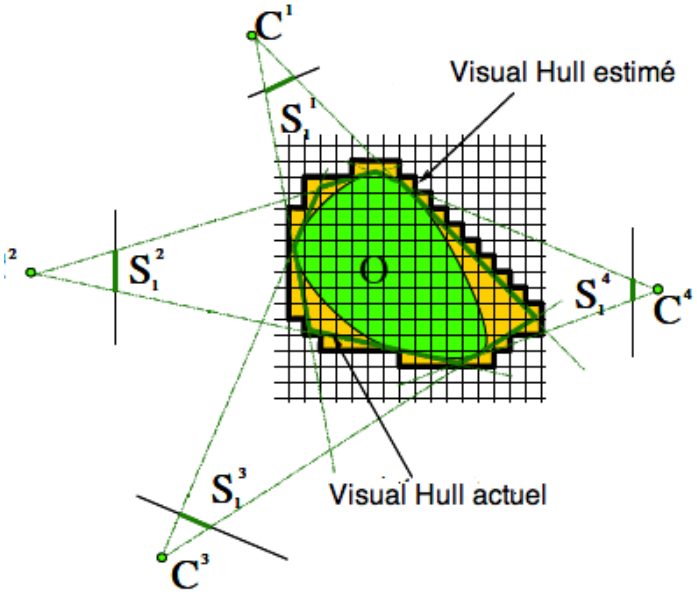


Figure 2.7 – Un exemple 2D d'une construction d'un *visual hull* par l'approche volumique. La région orange représente le *visual hull* estimé et la partie verte représente le vrai visual hull (Cheung, 2003)

Le pseudo-code de cette approche peut être donné comme ceci :

- Diviser l'espace de la scène en $N \times N \times N$ voxels discret X_n, $n = 1, ..., N^3$.
- Initialiser tous les voxels comme des voxels interne (appartient à l'objet) pour $n = 1, ..., N^3$.
- pour k=1 ... K, Projeter X_n dans l'espace image en utilisant P^k.
- Si $P^k X_n$ ne se projette pas dans l'espace de l'image silhouette alors X_n ne fait pas partie de la scène.
- Le *visual hull* H_j est estimé par l'union de tous les voxels.

La figure 2.7 illustre un exemple 2D de l'estimation de l'objet par l'approche Shape From Silhouettes volumique.

2.6 Algorithmes d'apprentissage machine

La classification est la dernière étape dans le processus de la reconnaissance de postures humaines. C'est principalement par apprentissage machine qu'on tente d'accomplir cette tâche. Dans cette optique, nous présentons différents algorithmes d'apprentissage machine qui ont été adaptés et appliqués dans notre recherche.

2.6.1 K-plus proches voisins

L'algorithme des k plus proches voisins (*k-nearest neighbours ou KNN*) est une technique d'apprentissage à base d'exemples. Elle ne comporte pas une phase d'entrainement en tant que telle. Elle cherche un groupe de k objets dans l'ensemble d'apprentissage, déjà emmagasinés, qui sont les plus proches de l'objet test. L'attribution d'une étiquette à l'objet test est faite en se basant sur la classe prédominante dans ce groupe de k objets.

Il existe trois éléments clés dans cette approche : 1) un ensemble d'objets étiquetés, 2) une métrique de similarité pour calculer la distance entre les objets 3) le nombre de plus proches voisins k. Pour classer un objet non étiqueté, les distances de cet objet aux objets étiquetés sont calculées, ses k plus proches voisins sont identifiés et les étiquettes de classes de ces voisins sont ensuite utilisées pour déterminer le libellé de classe de l'objet.

Un pseudo-code de l'algorithme KNN peut s'écrire comme suit :

Entrée : D, l'ensemble d'apprentissage de N exemples $(\mathbf{x}_i, y_i)$ avec $i = 1 \ldots N$ où $\mathbf{x}_i$ est un vecteur de caractéristique et y_i est son étiquette de classe correspondante, un objet de test $z = (\mathbf{x'}, y')$.

Processus :

- calculer $d(\mathbf{x'}, \mathbf{x})$, la distance entre z et chaque exemple $(\mathbf{x}, y)$ dans D.
- Sélectionner $D_z \subseteq D$, l'ensemble des objets d'entraînement les plus proches de z .

Sortie :

$$y' = \arg\max_{v} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v == y_i) \tag{2.11}$$

$$I(a == b) = \begin{cases} 1 \text{ si } a = b \\ 0 \textit{ sinon} \end{cases} \tag{2.12}$$

Différents problèmes peuvent affecter la performance de KNN :

- le choix de k : si k est trop petit, le résultat peut être sensible au bruit. D'autre part, si k est trop grand, alors le voisinage peut inclure trop de points d'autres classes.
- la technique de combinaison des étiquettes de K plus proches voisins. La méthode la plus simple est de prendre un vote majoritaire, mais cela peut poser un problème si les voisins les plus proches varient largement dans leur distance. Une approche plus sophistiquée, qui est généralement beaucoup moins sensible au choix de k, pondère le vote de chaque objet par sa distance, où le facteur de pondération est souvent considéré comme étant l'inverse de la distance au carré : $w_i = \frac{1}{d(\mathbf{x},\mathbf{x'})^2}$. Alors, l'équation 2.11 est remplacé par :

$$y' = \arg\max_{v} \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i I(v = y_i) \tag{2.13}$$

- le choix de la distance ; bien que différentes métriques puissent être utilisées pour calculer la distance entre deux points, la mesure de distance la plus souhaitable est celle pour laquelle une plus petite distance entre deux objets implique une plus grande probabilité d'avoir la même classe. Par exemple, certaines distances peuvent être affectées par la grande dimensionnalité des données. En particulier, il est très connu que la distance euclidienne devient moins discriminante lorsque le nombre des attributs devient grand.

La classification KNN est une technique de classification facile à comprendre et facile à mettre en œuvre. Malgré sa simplicité, elle peut bien fonctionner dans de nombreuses situations. En particulier, (Cover et Hart, 1967) montre que la probabilité d'er-

reur de KNN est majorée de deux fois par la probabilité d'erreur de Bayes sous certaines hypothèses raisonnables.

2.6.2 Machines à vecteurs de support

Les machines à vecteurs de support (en anglais *Support Vector Machines, SVM*) forment un ensemble de techniques d'apprentissage supervisé destinées initialement à résoudre des problèmes de classification binaire et de régression (Cortes et Vapnik, 1995). Ensuite, il a été adapté à la classification multiclasses. Le principe de SVM est de trouver l'hyperplan à vaste marge qui sépare correctement les données d'apprentissage. Cette marge est définie comme la distance du point le plus proche de l'hyperplan. Dans ce qui suit, tout d'abord, nous présentons la classification binaire des données linéairement séparables. Ensuite, nous détaillons la classification binaire de données non linéaire ainsi que l'extension de SVM à la classification multiclasses.

Classification linéaire

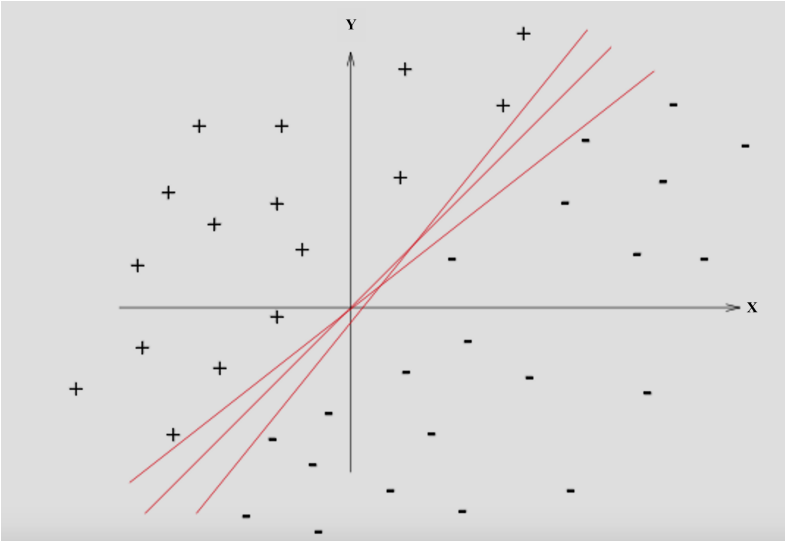


Figure 2.8 – Infinité d'hyperplans séparateurs pour des données linéairement séparables (Wikipedia, 2016b)

C'est le cas le plus simple où les données d'apprentissage viennent uniquement des

deux classes différentes +1 ou -1, c'est-à-dire que la classification est binaire. L'objectif de SVM est de rechercher un hyperplan qui sépare le mieux ces données. L'hyperplan peut être défini par l'équation suivante :

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{2.14}$$

où $\mathbf{w} \in \mathbb{R}^d$ est un vecteur de poids et $\mathbf{x} \in \mathbb{R}^d$ est un vecteur de données appartient à la base d'apprentissage $(\mathbf{x}_i, y_i)$ dont $y = +1$ ou $y = -1$ et $b \in \mathbb{R}$ est un paramètre. Pour inférer l'étiquette y d'un nouvel exemple du test $\mathbf{x}$, on utilise la fonction de décision suivante :

$$g(x) = \begin{cases} +1 \text{ si } f(x) > 0 \\ -1 \text{ si } f(x) < 0 \end{cases} \tag{2.15}$$

Si les données d'apprentissage $(\mathbf{x}', y)$ sont parfaitement linéairement séparables, c'est à dire ne contiennent pas des données bruitées (mal-étiquetées), alors il existe une infinité d'hyperplans qui séparent correctement ces données (voir figure 2.8). L'idée de SVM donc est de rechercher l'hyperplan optimal (celui qui maximise la marge) tout en garantissant que $y_i\mathbf{w}^T\mathbf{x}_i + b > 1$.

Pour déterminer l'hyperplan séparateur optimal, il nous faut tout d'abord calculer explicitement la distance euclidienne entre l'hyperplan et l'exemple le plus proche des deux classes (c.-à-d. la marge). Puisque le vecteur de poids $\mathbf{w}$ est orthogonal sur l'hyperplan optimal, alors la distance normalisée entre un exemple $\mathbf{x}$ et l'hyperplan est donnée par $\frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$ et par conséquent la marge normalisée est égale à $\frac{2}{\|\mathbf{w}\|}$ (voir la figure 2.9) Finalement, l'hyperplan séparateur optimal peut être obtenu en résolvant le problème d'équation :

$$\underset{\mathbf{w},b}{\operatorname{argmax}} \left\{ \frac{1}{\|\mathbf{w}\|} \min_i \left[y_i h(\mathbf{x}_i) \right] \right\} \tag{2.16}$$

Puisque les données sont censées être idéalement linéairement séparables, ce cas est appelé **SVM à marge dure (Hard marge)**.

En réalité, il n'est pas toujours possible de trouver une séparation linéaire puisque les données peuvent être bruitées. Il se peut aussi que la base d'apprentissage contienne

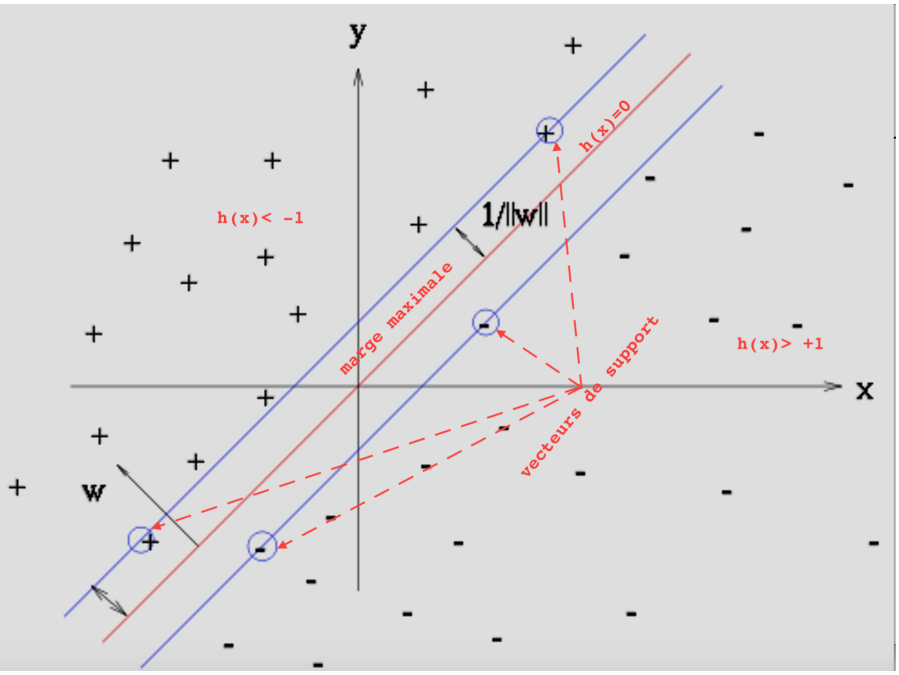


Figure 2.9 – Hyperplan séparateur optimal avec la marge maximale(adaptée de (Wikipedia, 2016b))

des erreurs d'étiquetage (*outiliers*) et que l'hyperplan séparateur ne soit pas la meilleure solution.

Dans le cas où les données d'apprentissage sont bruitées ou contiennent des erreurs, c'est-à-dire qu'elles ne sont pas linéairement séparables, l'équation $(y_i \mathbf{w}^T \mathbf{x})$ ne sera pas vérifiée pour tous les x_i ce qui nécessite une relaxation. En 1995, (Cortes et Vapnik, 1995) proposent une approche dite **SVM à marge souple** pour tolérer les mauvais classements. Cette approche consiste à chercher un hyperplan séparateur en minimisant le nombre d'erreurs de classification en introduisant des variables dites "ressorts" ou (*slack variables*) et généralement dénotées ξ_i. Par conséquent, l'équation 2.16 est réécrite comme :

$$\text{Minimiser } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i,\ C>0\ ,\ \xi_i \geq 0 \text{ et pour tout } i, y_i(w^T x_i + b) \geq 1-\xi_i \quad (2.17)$$

C est un paramètre permettant d'ajuster un compromis entre le nombre d'erreurs de

classification et la largeur de la marge.

Classification binaire non linéaire

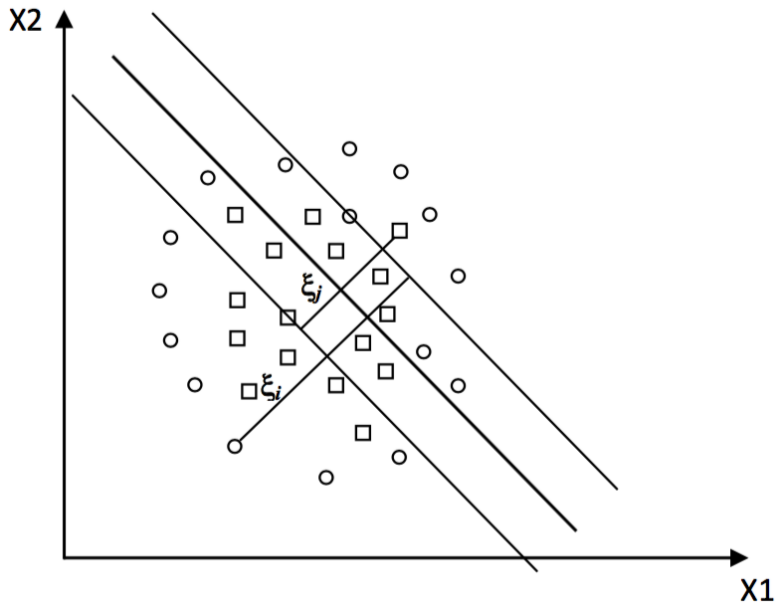


Figure 2.10 – Mal-adaptation de l'hyperplan à un problème non linéaire (adaptée de (Wikipedia, 2016b)).

En réalité les problèmes de classification sont souvent non linéaires. Alors, le fait de tolérer la mal-classification de certains exemples ne peut pas garantir toujours une meilleure généralisation pour un hyperplan. Par exemple, la figure 2.10 présente un problème où l'adaptation d'un séparateur linéaire ne convient pas pour une bonne généralisation. Cependant, ce problème peut être résolu par l'ajustement d'une fonction de discrimination non linéaire. L'ajustement d'une telle fonction est très difficile, voire impossible. Pour remédier à ce problème, l'idée est de projeter les données dans un espace d'une dimensionnalité plus grande (appelé espace de caractéristiques) où les données deviennent linéairement séparables (voir la figure 2.11). Cette projection est faite à travers

une fonction, qui peut être définie par :

$$\begin{aligned} &\mathbb{R}^d \mapsto \mathbb{R}^{d'} \\ &\mathbf{x'} = \phi(\mathbf{x}) \end{aligned} \tag{2.18}$$

où $d <<< d'$.

En pratique, la transformation ϕ est inconnue et on construit plutôt une fonction noyau qui respectent certaines conditions. Une fonction noyau doit correspondre à un produit scalaire dans l'espace de caractéristiques.

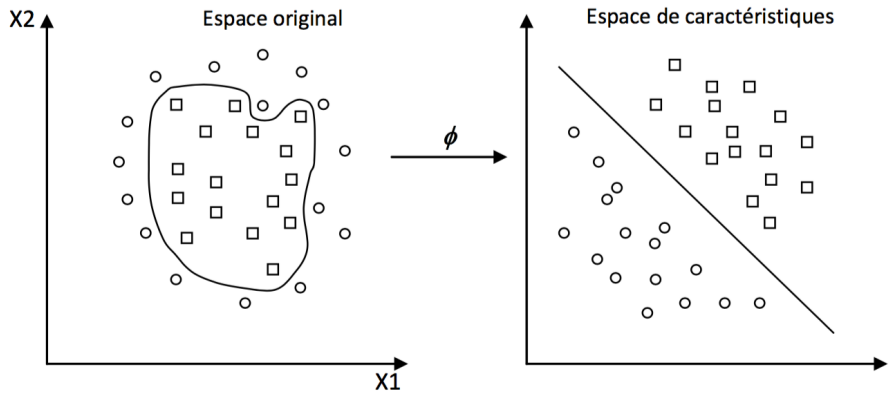


Figure 2.11 – Transformation d'espace par le noyau ϕ

Il existe certaines fonctions noyau usuelles utlisées avec SVM, par exemple :

- le noyau polynomial $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d$.
- le nayau gaussien $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right\}$.

Classification multi-classes

Les machines à vecteurs de support (Cortes et Vapnik, 1995) ont été conçues à l'origine pour la classification binaire. Cependant, les problèmes du monde réel sont généralement multiclasses, un exemple simple est la reconnaissance de postures humaines. Dans ce cas, la fonction de décision n'est plus binaire, c'est-à-dire qu'on ne cherche pas à affecter un nouvel exemple à l'une des deux classes, mais plutôt à l'une de plusieurs classes. Plusieurs méthodes ont été proposées pour étendre le SVM pour la classification

multiclasses (Abe, Hamel, 2011). Ces méthodes consistent à transformer les exemples d'apprentissage en plusieurs sous-ensembles représentant chacun un problème de classification binaire, qui peut être résolu par le SVM binaire. On trouve dans la littérature deux grandes approches de décomposition :

Un-contre-tous

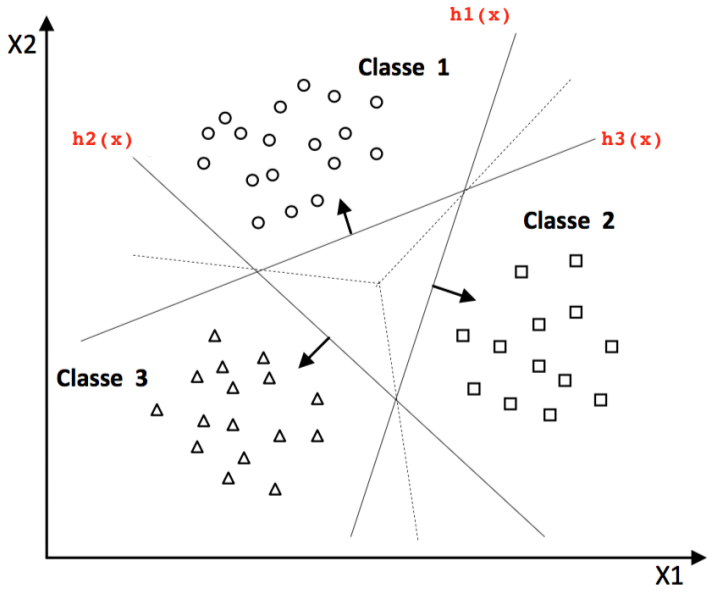


Figure 2.12 – Approche un-contre-tous

Cette approche consiste à construire k modèles SVM où k est le nombre des classes. L'apprentissage du $k^{\text{ème}}$ SVM est fait en supposant que tous les exemples de la $k^{\text{ème}}$ classe sont étiquetés avec étiquettes positives, et tous les autres exemples sont étiquetés avec des étiquettes négatives.

Pour classifier un nouvel exemple **x**, l'étiquette y est obtenue en appliquant le principe "winner-takes-all" : l'étiquette affectée à **x** est celle associée au classifieur ayant renvoyé le score le plus élevé.

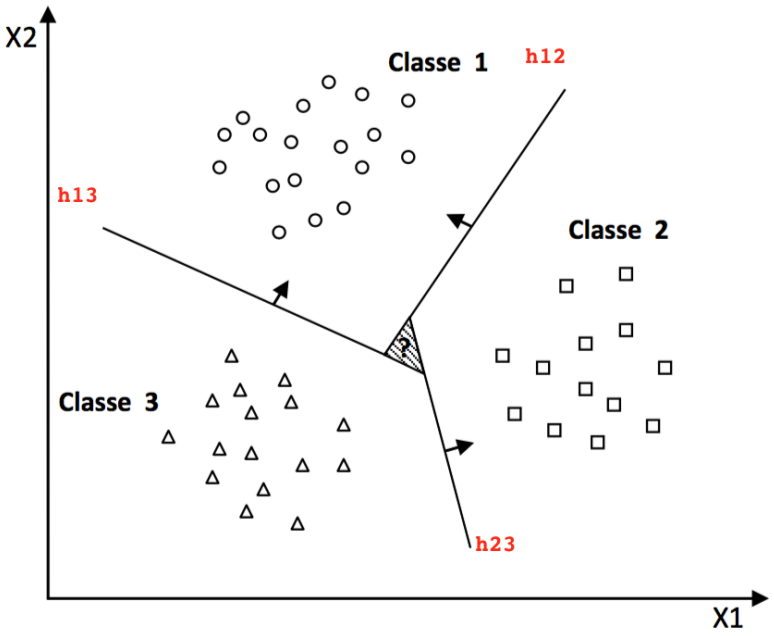


Figure 2.13 – Approche un-contre-un

Un-contre-un

Cette approche est aussi appelée “pairwise” consiste à construire un classifieur pour chaque paire de classes. Par conséquent, elle apprend $\frac{k(k-1)}{2}$ fonctions de décisions pour k classes. L’affectation d’un nouvel exemple est faite par vote majoritaire. On évalue tous les $\frac{k(k-1)}{2}$ classifieurs, et pour chaque test, on vote pour la classe à laquelle appartient l’exemple. Finalement, le nouvel exemple est affecté à la classe la plus votée.

2.6.3 Eigenpostures par décomposition en valeurs singulières

L’idée des "eigenpostures" est inspirée à partir de la technique des “eigenfaces”, qui a été utilisée par (Turk et Pentland, 1991) pour la reconnaissance de visages. Les “eigenpostures” sont des vecteurs propres utilisés pour représenter le sous-espace des postures avec une base orthonormale pour modéliser les variations d’une posture. Dans ce manuscrit, nous avons utilisé la décomposition en valeurs singulières (SVD) pour construire une telle base orthogonale.

Étant donné N images de taille $m = n \times n$ (dans notre cas), qui représentent une

seule classe de postures (par exemple, "debout"), on définit une matrice A de dimension $m \times N$ où chaque colonne représente une image vectorisée, c.-à-d. que toutes les colonnes d'une image sont empilées l'une après l'autre. La matrice A définit un sous-espace d'une posture dans $\mathbb{R}^m$. Cependant, ce sous-espace a généralement une dimension très petite, sinon les différents sous-espaces de classes (par exemple, classes de postures) se chevaucheraient et on ne pourrait pas les distinguer ! On utilise les techniques de la réduction de la dimensionnalité afin de réduire la dimension du sous-espace tout en garantissant qu'il modélise parfaitement les données de la classe. Une méthode très utilisée dans ce contexte est la décomposition en valeurs singulières (SVD).

Décomposition en valeurs singulières et construction d'une base orthonormale

$$A = U\Sigma V^T \tag{2.19}$$

$$= \Big(\underbrace{u_1 \ \cdots \ u_r}_{\text{col}(A)} \ \underbrace{u_{r+1} \ \cdots \ u_{m=n^2}}_{\text{null}(A)} \Big) \begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_r^T \\ v_{r+1}^T \\ \vdots \\ v_N^T \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{row}(A) \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{null}(A) \end{matrix} \tag{2.20}$$

- la matrice U contient un ensemble de vecteurs de base orthonormés.
- la matrice Σ contient les valeurs singulières dans ses coefficients diagonaux de la matrice A où les σ_i sont ordonnées par ordre décroissant.
- la matrice V contient un ensemble de vecteurs de base orthonormés.

Chaque colonne a_j de A représente une posture d'une même classe et peut s'écrire :

$$a_j = \sum_{i=1}^{r} (\sigma_i v_{ji}) u_i \tag{2.21}$$

D'après l'équation 2.21, chaque posture (colonne) est une combinaison linéaire de vecteurs (images) singuliers u_i qui forment une base orthonormale. Pour alléger l'écriture, on réécrit l'équation 2.21 :

$$a = \sum_i \alpha_i u_i \tag{2.22}$$

Pour réduire la dimensionnalité, on va tronquer la SVD en ne conservant que les r plus grandes valeurs singulières et vecteurs singuliers associés.

Algorithme de classification

- Soit z une nouvelle posture.
- Étant donnés les sous espaces $U_{1r}, U_{2r}, \ldots, U_{Cr}$ où C dénote le nombre de classes et $U_{cr} = (u_{c1}, u_{c2}, \ldots, u_{cr})$ où $c = 1 \ldots C$
- Projeter z dans les différents sous espaces $U_c r$ et calculons les résidus correspondants :

$$res_c = \min_{\alpha_{ci}} \left\| z - \sum_{i=1}^{r} \alpha_{ci} u_{ci} \right\|_2 \tag{2.23}$$

- z est affecté à la classe qui renvoit le résidu le plus faible.

Supposons que $\sum_{i=1}^{r} \alpha_{ci} u_{ci} = U_{cr} \alpha_c$. Puisque U_{cr} est orthonormale, les équations normales nous donnent la solution :

$$\alpha = U_{cr}^T \cdot z \tag{2.24}$$

Le résidu donné par l'équation 2.23 est alors donné tout simplement par :

$$\left\| (I - U_{cr} U_{cr}^T) \cdot z \right\|_2 \tag{2.25}$$

2.6.4 Perceptron multicouches (MLP) et réseau de neuronnes convolutionnel (*CNN*)

Perceptron multicouches

Un neuronne formel est une modélisation mathématique permettant de simuler le fonctionnement du neurone biologique, en particulier la sommation des entrées. Donc, un neuronne (voir la figure 2.14) peut être considérée comme une fonction mathématique

réelle à plusieurs variables $x_1, x_2, \ldots, x_m$ pondérées respectivement par différents poids $w_1, w_2, \ldots, w_m$. Dans la première modélisation de McCulloch et Pitts, un neurone calcule la somme de ses entrées pondérées $w_1x_1 + w_2x_2 + \cdots + w_mx_m$ et la compare à son seuil w_0. Si le résultat est supérieur au seuil alors la sortie y associée aux variables est 1, sinon elle vaut 0. D'où la formule 2.26, où f est la fonction d'activation (fonction de *Heaviside* dans le cas McCulloch et Pitts).

$$y = f(\mathbf{x}) = \begin{cases} 1 \; si \; w_1x_1 + w_2x_2 + ... + w_mx_m > w_0. \\ 0 \; sinon. \end{cases} \tag{2.26}$$

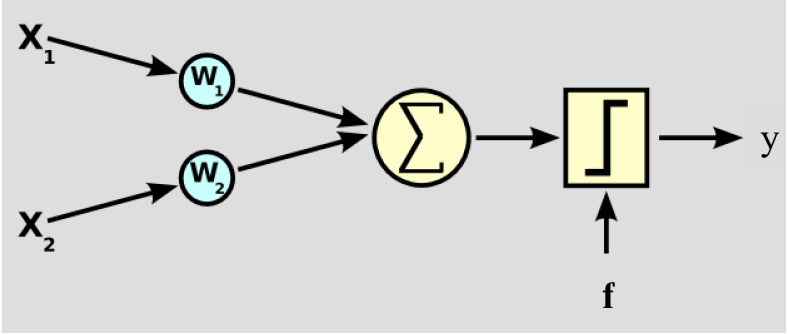


Figure 2.14 – Neurone formel avec 2 entrées et une fonction d'activation à seuil (McCulloch et Pitts).

Fonctions d'activations : Depuis la première modélisation, différentes fonctions d'activation ont été proposées, par exemple :

Sigmoide : $f(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}}$

ReLU : $f(\mathbf{x}) = max(0, \mathbf{x})$

Leaky ReLU : $f(\mathbf{x}) = max(a\mathbf{x}, \mathbf{x})$

Maxout : $f(\mathbf{x}) = max(a_0\mathbf{x} + b0, a_1\mathbf{x} + b_1)$.

Le réseau de neurones multicouches (MLP) est un classifieur organisé en plusieurs couches successives où l'information circule uniquement de la couche d'entrée vers la couche de sortie et n'est jamais retournée en arrière ; il s'agit d'un réseau de neurones à

propagation avant (*feedforward*). Un MLP consiste en une couche d'entrée , une couche de sortie et une ou plusieurs couches cachées (voir la figure 2.15). Chaque couche possède un nombre variable de neurones sauf la couche de sortie qui contient toujours un nombre de neurones égal aux sorties du système (par exemple, le nombre de classes dans le cas de classification). Un réseau de neurones multicouches consiste en une si-

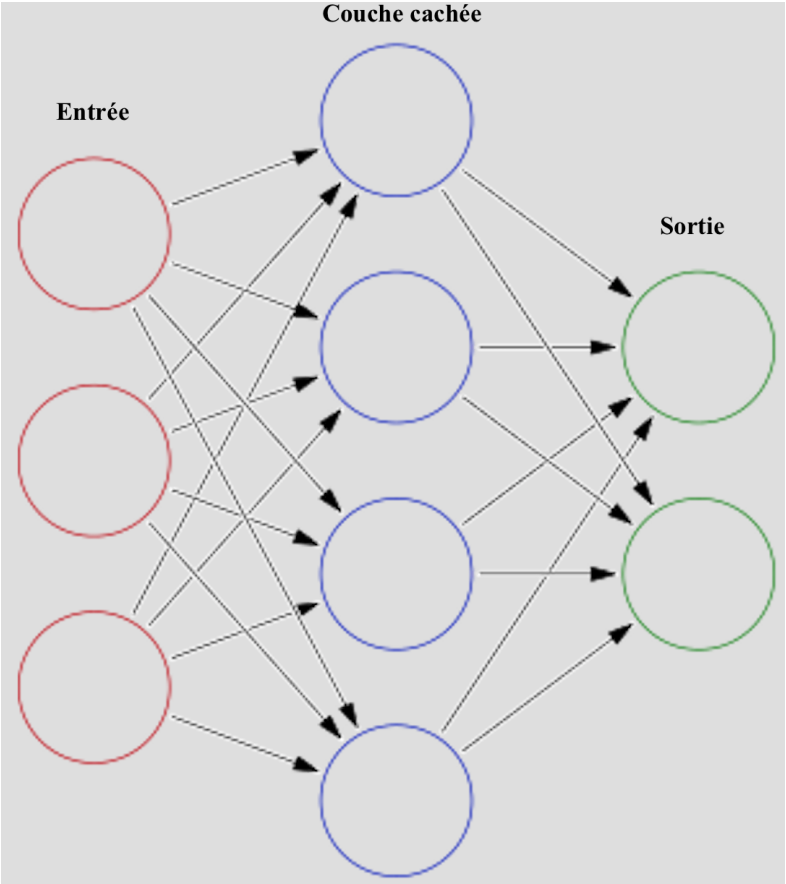


Figure 2.15 – Perceptron multicouches.

mulation d'une fonction $y = f(\mathbf{x},\mathbf{w})$ en minimisant une fonction de coût qui permet de mesurer la façon dont la prédiction $\hat{y}$ correspond à la valeur réelle y. Les fonctions de coût couramment utilisées sont :

fonction carré : $(y-\hat{y})^2$

fonction cross-entropy : $-\sum_i y_i log(\hat{y}_i)$

avec y_i est l'étiquette de la classe (0 ou 1 dans le cas de classification binaire) et $\hat{y}_i$ est le prédicteur correspond à une probabilité de classe.

Pendant l'apprentissage d'un MLP, on tente de minimiser la perte totale sur un ensemble des données d'apprentissage. En fait, on veut trouver le vecteur du poids w^* qui minimise le coût total :

$$w^* = \text{argmin} \sum_i coût(f(x_i, w), y_i) \tag{2.27}$$

Malgré la puissance des MLPs, dont un réseau de neurones avec une seule couche cachée ayant un nombre fini de neurones est capable d'approximer n'importe qu'elle fonction continue sous certaines hypothèses légères sur la fonction d'activation (Hornik, 1991), l'utilisation de couches entièrement connectées présente un grand problème de calcul lors du traitement d'images. Par exemple, pour une image de taille 10×10, un MLP de 3 couches avec 200 neurones chacune contient ~100k paramètres. En se basant sur cette observation : Les MLPs peuvent être améliorés de deux façons :

- Utilisation de couches localement connectées au lieu de couches entièrement connectées.
- Utilisation des poids partagés entre les neurones.

Ces améliorations ont été réalisées dans les réseaux de neurones convolutionnels.

Réseaux de neuronnes convolutionnels

Les réseaux de neurones convolutionnels (*convolution neural network ou CNN*) peuvent être vus comme une combinaison un peu étrange de la biologie et des mathématiques, mais ces réseaux ont été parmi les innovations les plus influentes dans le domaine de la vision par ordinateur. Ils ont été originalement proposés par (LeCun et al., 1998) en 1998. Cependant, les réseaux convolutionnels ont gagné beaucoup d'importance et sont devenus populaires après leurs utilisations par le chercheur (Krizhevsky et al., 2012) pour gagner le concours ImageNet de l'année 2012, ce qui a fait passer l'erreur de classification de 26% à 15%, une amélioration étonnante à l'époque. Depuis, différentes

entreprises ont utilisé l'apprentissage profond au coeur de leurs services. Facebook utilise des réseaux neuronaux pour leurs algorithmes de marquage automatique, Google pour leur recherche de photo, Amazon pour leurs recommandations de produits, etc.

Cependant, le cas d'utilisation classique, et sans doute le plus populaire, de ces réseaux est pour le traitement d'image et, par conséquent, le traitement de vidéo. Dans ce qui suit, nous présentons brièvement l'architecture d'un réseau de neurones convolutionnel, quelques techniques utilisées pour l'optimisation d'hyperparamètres et des stratégies de régularisation.

Architecture d'un réseau convolutionnel

Les réseaux de neurones convolutionnels sont conçus pour traiter une image bidimensionnelle (2D), tridimensionnelle(3D), voire quadridimensionnelle (4D). Un CNN se compose de trois principaux types de couches : (i) des couches convolutionnels (ii) des couches de sous-échantillonnage (*pooling*), et (iii) une couche de sortie (couche entièrement connectée). Les couches de réseau sont organisées dans une structure empilée : chaque couche de convolution est suivie d'une couche de *pooling* et la dernière couche de convolution est suivie par la couche de sortie (voir figure 2.16). Les couches de convolution et de *pooling* sont considérées comme des couches 2D, tandis que la couche de sortie est considérée comme une couche 1D. Dans un réseau convolutionnel, une couche 2D a plusieurs plans. Un plan se compose de neurones qui sont stockés dans un tableau bidimensionnel. La sortie d'un plan est appelée une carte de caractéristiques (*feature maps* ou f.m). Par conséquent, chaque plan de *pooling* réduit sa taille d'entrée de moitié, le long de chaque dimension. Une carte de caractéristiques dans une couche de sous-échantillonnage est connectée à un ou plusieurs plans dans la couche de convolution suivante.

- Dans une **couche convolutionnelle**, chaque plan est relié à une ou plusieurs cartes de caractéristiques de la couche précédente (par exemple, si l'entrée est une image couleur, chaque plan est relié à 3 cartes de caractéristiques.). Une connexion est associée à un masque de convolution, qui est une matrice 2D (filtre)

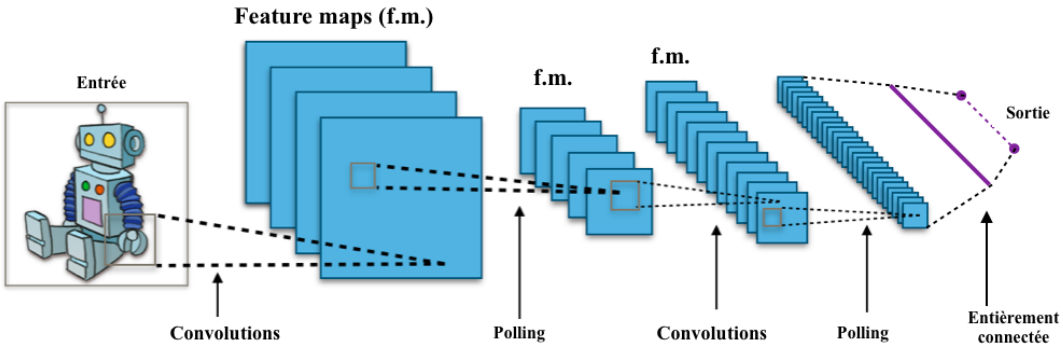


Figure 2.16 – Architecture d'un *CNN* (Wikipedia, 2016a).

d'entrées ajustables (pendant la phase d'entrainement) appelées poids partagés. Chaque plan applique une convolution entre ses entrées 2D et ses masques de convolution. Les sorties de convolution sont additionnées et ensuite ajoutées à un scalaire ajustable, noté terme de *bias*. Finalement, une fonction d'activation est appliquée sur le résultat pour obtenir la sortie du plan. La sortie du plan est une matrice 2D appelée cartes de caractéristiques. Telle que son nom indique, chaque carte de caractéristiques indique la présence d'une caractéristique (*feature*) à un emplacement d'un pixel donné. Une couche de convolution exhibe une ou plusieurs cartes de caractéristiques. Chacune est ensuite connectée à exactement un plan dans la couche de *pooling* suivante.

- **Une couche de *pooling*** a le même nombre de plans que la couche de convolution précédente. Chaque plan dans la couche de *pooling* divise son entrée 2D en blocs non chevauchants de taille, en général, 2×2 pixels. Par conséquent, chaque plan de *pooling* réduit sa taille d'entrée de moitié, le long de chaque dimension. Donc, elle réduit le nombre de paramètres dans le réseau et peut contrôler le surapprentissage. En plus, une couche de *pooling* crée aussi une forme d'invariance par translation. Enfin, une carte de caractéristiques dans une couche de *pooling* est connectée à un ou plusieurs plans dans la couche de convolution suivante.
- **La couche de sortie**, en général, peut être construite à partir d'une fonction régression logistique. Cette fonction donne un vecteur numérique représentant une distribution de probabilité (sommé à 1) qui indique la probabilité d'appartenance

d'un exemple **x** aux différentes classes. La classe avec la plus grande probabilité est affectée à **x**.

Optimisation d'hyperparamètres

Les hyperparamètres signifient en fait les paramètres qui influencent d'une certaine façon sur l'architecture d'un CNN ainsi que le temps nécessaire pour son apprentissage. Le taux d'apprentissage, le nombre de couches cachées ainsi que le nombre de neurones cachés et les constantes de régularisation sont considérés parmi les hyperparamètres dans le contexte d'un réseau de neurones multicouches (MLP). Cependant, les CNNs ont beaucoup plus d'hyperparamètres qu'il faut prendre en considération tels que : le nombre de couches convolutionnelles, le nombre et la taille des filtres. Un autre hyperparamètre est le type de *pooling* (max-pooling, mean-pooling, sum-pooling, etc).

L'optimisation des hyperparamètres est le problème du choix d'un ensemble d'hyperparamètres pour un algorithme d'apprentissage, généralement dans le but d'optimiser une mesure de la performance de l'algorithme sur un ensemble de données indépendant de l'ensemble de test. Différentes méthodes ont été proposées dans la littérature pour l'optimisation des hyperparamètres (Bengio, 2012).

Sélection par grille ou *Grid search* : Elle est effectuée en spécifiant tout simplement une liste de valeurs pour chaque paramètre et en essayant toutes les combinaisons possibles de ces valeurs. Cela peut paraître méthodique et exhaustif. La recherche par grille doit être évaluée par une métrique de performance, typiquement une validation croisée (*croos-validation*) sur les données d'apprentissage ou en utilisant la technique de *leave-one-out*. Cette stratégie devient potentiellement très coûteuse avec l'augmentation de la taille de listes de valeurs spécifiées.

Sélection aléatoire ou *Random search* : proposé par (Bergstra et Bengio, 2012), c'est le même principe que la recherche par grille, sauf qu'au lieu d'utiliser une recherche exhaustive, cette approche consiste à sélectionner les valeurs de paramètres aléatoirement à partir des listes définies, puis répète ce processus un nombre fixe d'itérations. Un modèle est construit et évalué pour chaque combi-

naison de paramètres choisis. La recherche aléatoire des paramètres est garantie d'être plus efficace que la recherche par grille dans des espaces de grande dimension. En effet, il se trouve que certains hyperparamètres n'affectent pas de façon significative la fonction de coût, et par conséquent, choisir aléatoirement un ensemble de valeurs dispersées est jugé plus efficace que la recherche exhaustive dans des paramètres qui n'ont finalement pas d'influence sur la fonction de coût.

Régularisation

La régularisation est une technique très importante dans l'apprentissage automatique pour éviter le surapprentissage. Mathématiquement, elle ajoute un terme de régularisation pour éviter que les coefficients ne s'adaptent parfaitement. Dans la littérature, il existe plusieurs façons de contrôler la capacité des réseaux de neurones afin de prévenir le problème de surapprentissage et améliorer la généralisation.

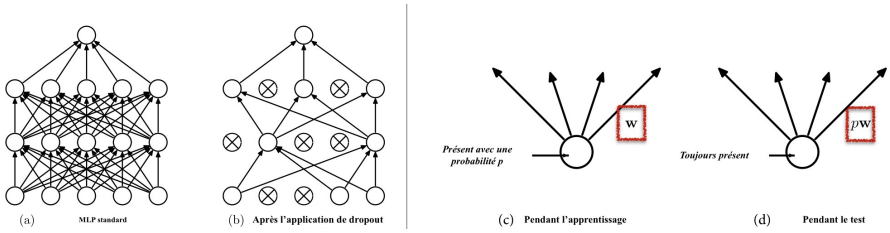


Figure 2.17 – Technique de dropout (Srivastava et al., 2014b)

La régularisation L1 ajoute juste la somme des poids. Alors la fonction objectif 2.27 se réécrit comme suit :

$$w^* = \text{argmin} \sum_i coût(f(x_i, w), y_i) + \lambda \sum_k^m |w_k| \tag{2.28}$$

La régularisation L2 ajoute la somme du carré des poids à la fonction du coût. Par conséquent, l'équation 2.27 devient :

$$w^* = \text{argmin} \sum_i coût(f(x_i, w), y_i) + \lambda \sum_k^m w_k^2 \tag{2.29}$$

Dropout c'est une technique de régularisation extrêmement efficace, simple et récemment introduite par (Srivastava et al., 2014b). Pendant l'apprentissage, le Dropout peut être interprété comme l'échantillonnage d'un réseau de neurones, avec une probabilité p, au sein du réseau de neurones complet, où la mise à jour des poids est faite seulement pour le réseau échantillonné sur la base des données d'apprentissage. Pendant la phase de test, on garde tous les neurones actifs et la prédiction renvoyée peut être interprétée comme une prédiction moyenne à travers tous les sous-réseaux échantillonnés dans la phase d'apprentissage (voir la figure 2.17).

2.7 Utilisations des ombres

Les applications de vision par ordinateur intègrent souvent des ombres dans leurs modèles vu que leurs présences dans une image fournissent des informations puissantes pouvant être utilisées pour déterminer la forme 3D et les orientations des objets dans la scène. L'interprétation des ombres dans une image implique différentes considérations techniques notamment le type d'éclairage (soleil, infrarouge, visible), le nombre de sources lumineuses utilisées, la façon d'extraire l'ombre et de gérer l'interférence entre les différentes sources, l'effet de l'éclairage ambiant, etc. Les premiers travaux ont porté sur l'interprétation des ombres à partir des dessins au trait (*line drawings*) : (Shafer et Kanade, 1983) ont établi des contraintes fondamentales permettant de récupérer l'orientation de surfaces à partir de l'observation de l'ombre projetée. Plus tard, Bouguet et Perona (1999) ont présenté un système simple et peu coûteux formé par une caméra, une source lumineuse visible, un bâton et un damier pour extraire la forme tridimensionnelle des objets. Ce système fonctionne de façon similaire à la technique de la lumière structurée. En déplaçant le bâton devant la source lumineuse, une ombre mobile se projette sur la scène. La forme 3D de l'objet est extraite de l'emplacement spatial et temporel de l'ombre observée. (Yamazaki et al., 2007), ont proposé une nouvelle approche appelée imagerie de shadowgrams coplanaires (*Shadow grams coplanar imaging*). Celle-ci est illustrée dans la figure 2.18. En translatant seulement la source lu-

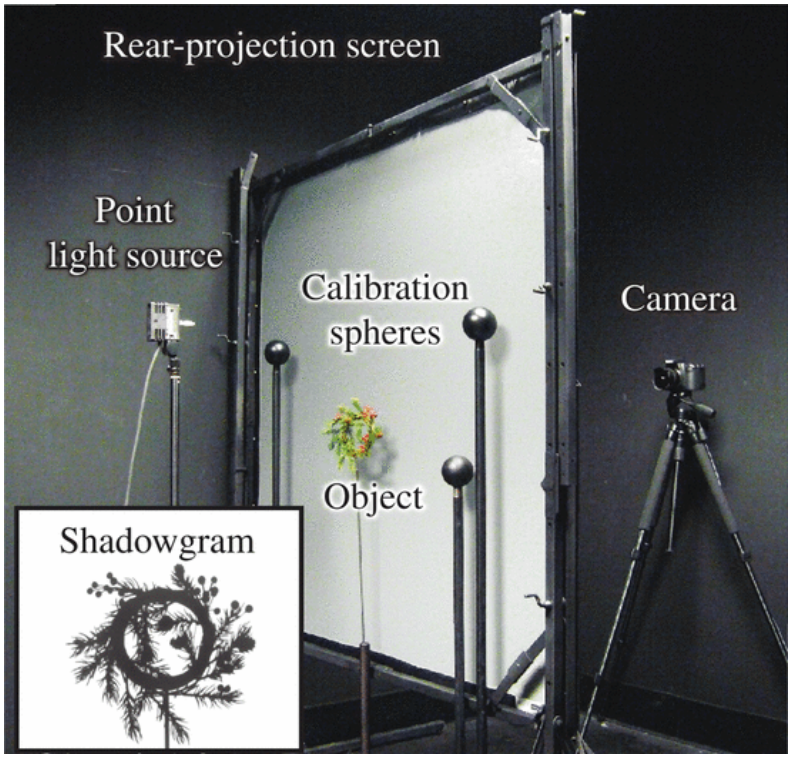


Figure 2.18 – Le système utilisé pour capturer les "*shadowgrams*" coplanaires comprend une caméra numérique, une source lumineuse à un seul point et un écran de projection arrière. L'objet est placé à proximité de l'écran pour couvrir un grand champ de vision. Deux ou plusieurs sphères sont utilisées pour estimer les positions initiales de la source lumineuse. (image insérée) Un exemple de *"shadowgram"* obtenu à l'aide de cette configuration (Yamazaki et al., 2007).

mineuse, différente shadowgrams (silhouettes d'ombres) coplanaires sont obtenues. La forme 3D de l'objet (*visual hull*) est obtenue en développant une nouvelle formulation plus simple de la géométrie épipolaire. (Savarese et al., 2007) a proposé un système très similaire à celui de (Yamazaki et al., 2007), la seule différence est de tourner l'objet au lieu de déplacer la source lumineuse à fin d'obtenir différentes shadowgrams. Dans les travaux cités ci-dessus, des dispositifs très bien contrôlés ont été utilisés pour éviter certains problèmes techniques. En fait, une seule source lumineuse est utilisée, l'objet est supposé fixe et l'arrière-plan est une surface plane et ne contient que la silhouette d'ombre. En outre, à chaque fois une seule silhouette d'ombre par image est obtenue afin d'éviter l'interférence entre les ombres.

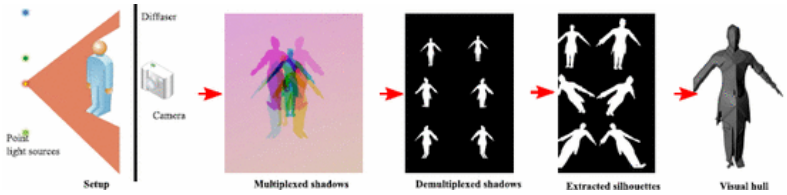


Figure 2.19 – Le système se compose de plusieurs sources lumineuses ponctuelles colorées, d'un diffuseur et d'une caméra numérique. Pour chaque source lumineuse, une silhouette est extraite des ombres capturées de la scène (Cuypers et al., 2009)

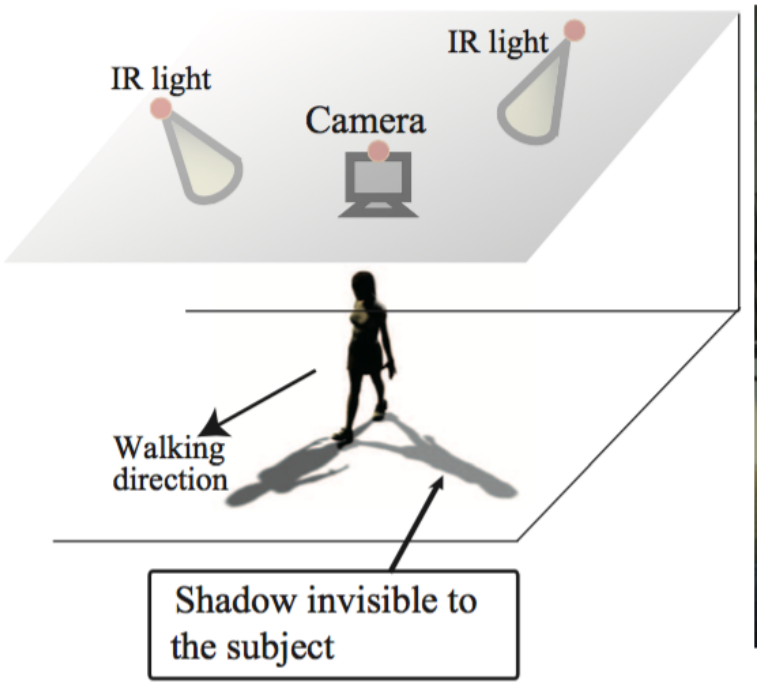


Figure 2.20 – Identification de la démarche à l'aide d'ombres invisibles (Iwashita et al., 2013)

(Cuypers et al., 2009) ont proposé un système multivues en utilisant une seule caméra et un ensemble de sources lumineuses ponctuelles pour récupérer le *visual hull* d'un objet (voir figure 2.19). Au contraire des travaux précédents, plusieurs sources lumineuses ont été utilisées et allumées au même temps. Cependant, les sources lumineuses sont supposées d'être de différentes couleurs afin d'éviter le problème d'interférence entre les différentes ombres et faciliter leurs extractions.

Dernièrement, (Iwashita et al., 2012a, 2013, 2012b) ont introduit un système mul-

tivues en utilisant une seule caméra et 2 sources lumineuses infrarouges (voir la figure 2.20). Ce système est utilisé pour l'identification de la démarche dont le but est la reconnaissance des personnes dans un contexte de sécurité dans un aéroport. Bien que ce système est installé dans un environnement moins contrôlé, plusieurs mesures ont été prises afin de faciliter l'extraction des ombres. Un filtre IR-bande est utilisé pour ne laisser passer que l'infrarouge afin de diminuer l'influence de la lumière ambiante sur la qualité de l'ombre. En outre, les sources lumineuses ont été placées obliquement des deux côtés de la personne qui est toujours debout ce qui empêche l'interférence des deux ombres et facilite considérablement l'extraction des ombres.

Dans cette thèse, nous proposons une approche similaire à celle de (Iwashita et al., 2013) dont le dispositif expérimental requiert plusieurs équipements matériels tels que :

- 2 sources lumineuses infrarouges ou plus avec 140 diodes chacune.
- Une caméra Prosilica GC1380 avec un filtre IR-bande et une lentille à grand angle (110 °).
- Un relais (*Relay Shield*) controlé par un microcontrôleur Arduino.
- Un tissu en polyester.
- Un grand miroir acrylique.
- Un laptop.

En utilisant ce dispositif, notre objectif est la reconnaissance de postures humaines dans un environnement intérieur. Dans un premier travail et pour valider notre approche, nous avons installé 4 sources lumineuses dans les coins supérieurs d'une chambre et une caméra avec un filtre IR-bande est placée dans le centre du plafond. Bien que l'utilisation de la lumière visible reste toujours possible, l'utilisation d'une lumière infrarouge dans notre cas est essentielle permettant au système de fonctionner jour et nuit sans déranger l'utilisateur. En allumant les sources lumineuses au même temps, la scène devient assez illuminée ce qui affaiblit gravement la netteté des ombres. En outre, les ombres peuvent se chevaucher et leur séparation devient difficile. Pour remédier à ces problèmes, les sources lumineuses ont été contrôlées par un dispositif électronique afin d'allumer cycliquement une source à la fois. Ceci nous permet d'éviter le chevauchement des ombres et d'avoir une seule ombre par image, et par conséquent, l'extraction des ombres devient

triviale. Cependant, un problème de synchronisation entre les différentes sources peut survenir si la personne bouge rapidement.

Dans des conditions de lumière visible, les vues de caméra sont semblables à ce que nous voyons avec les yeux. Mais, en utilisant de la lumière infrarouge, elles sont radicalement différentes. Les caractéristiques spectrales (absorptivité, réflectivité, transmissivité) d'infrarouge (IR) de différentes fibres et textiles sont fortement liées au type de fibre, à la densité de surface, à la récupération d'humidité et à la construction de tissu (McFarland et al., 1999). Dans notre cas, les images de caméra ont été très sombres à cause de la faible réflectivité des tuiles du sol, et ceci a affecté sérieusement la qualité des ombres projetées. Pour résoudre ce problème et en nous basant sur cette étude (Zhou et al., 2011b), nous avons utilisé un tissu en polyester ayant une grande réflectivité.

Dans des conditions de lumière invisible (infrarouge), l'éclairage ambiant peut influencer sur la qualité des ombres projetées même avec l'utilisation d' un filtre IR-bande (avec la caméra). En effet, la lumière du soleil contient de l'infrarouge qui peut passer à travers le filtre et générer une ombre supplémentaire. "En mode laboratoire", nous avons résolu ce problème en bloquant les fenêtres par des bandes noires. Toutefois, dans un contexte plus réel, des filtres IR-bande peuvent être installés sur les fenêtres pour empêcher l'effet de la lumière ambiante.

Dans nos travaux subséquents, nous avons supposé que la scène est vide et nous n'avons pas traité l'influence de la présence de meubles sur la performance de notre approche. Cependant, on s'attend que la présence de meubles dans la scène puisse causer de problèmes pour l'extraction des ombres projetées sur des différentes surfaces (meubles, mur, sol...). En outre, les meubles peuvent avoir des faibles réflectivités et nous devons utiliser des tissus en polyester qui va compliquer beaucoup plus la scène. Cependant, une solution qui semble possible est d'utiliser un réseau de neurones convolutionnel pour apprendre toute la scène.

CHAPITRE 3

HUMAN POSTURE RECOGNITION BY COMBINING SILHOUETTE AND INFRARED CAST SHADOWS (ARTICLE)

Ce chapitre présente le manuscrit intitulé "*Human posture recognition by combining silhouette and infrared cast shadows*" publié dans la conférence *Image processing theory, Tools and Applications* (IPTA 2015) par Rafik Gouiaa et Jean Meunier.

3.1 Avant-propos

Nous nous sommes intéressés à la reconnaissance de la posture humaine en nous basant sur la combinaison de l'information de silhouette et d'ombre invisible projetée par une lumière infrarouge capturées par une seule caméra. Différentes méthodes ont été proposées pour la reconnaissance de postures humaines en utilisant une seule caméra, par exemple (Kellokumpu et Heikkilä, 2005, Nadia et al., 2008) et (Girondel et al., 2005). Cependant, la performance de ces méthodes est souvent dépendante de point de vue de la caméra, des occultations et des ambiguïtés perspectives.

Nous avons donc proposé un système multivues basé sur une caméra et un nombre limité de sources lumineuses infrarouges pour la reconnaissance de postures humaines. Dans ce travail, notre montage expérimental consiste en 4 sources lumineuses infrarouges installées dans les coins supérieurs d'une chambre et une caméra avec un filtre infrarouge fixée dans le plafond pour capturer l'évolution du mouvement de la personne dans la scène. Les sources lumineuses sont allumées cycliquement une à la fois en utilisant un microcontrôleur Arduino afin d'obtenir une seule ombre par image (*frame*). En plus, pour étudier la faisabilité de cette approche, nous considérons une simple scène où l'ombre se projette directement sur le plancher. Notre méthode pour la classification de postures consiste à extraire une carte de distances (*distance transform*) à partir d'une image combinant la silhouette du corps avec l'ombre correspondante (4 cartes de distances), puis à appliquer un vote majoritaire afin d'inférer la posture en question. Notre

méthode a été testée sur des vidéos réalistes illustrant quelques postures normales et anormales enregistrés dans notre laboratoire.

Les notations utilisées dans cet article sont liées à l'article et n'ont pas de lien avec le reste de la thèse.

3.2 Abstract

This paper proposes a multi-infrared lights system for human posture recognition, which uses as input the combination of a body silhouette and cast shadows. More precisely, in our experimental setup, we installed 4 infrared lights in the different upper corners of a room to project shadows of a person, and a camera with an infrared filter in the ceiling to capture images. A simple electronic device is used to turn on and off each light in turn to get one cast shadow by frame. To illustrate the feasibility of this approach, we consider a simple scene where the shadows are projected directly on the ground. The features used for recognition are based on the distance transforms of the combination of the body silhouette and each cast shadow. A weighted majority vote scheme is used to decide what is the corresponding posture. Our approach was validated on a new data set captured in our laboratory, and compared with a traditional mono-camera system.

Keywords : Cast Shadows, Infrared illumination, Posture recognition, Distance transform, Activity recognition.

3.3 Introduction

This paper presents a new approach to recognize human postures in a controlled indoor environment such as for monitoring elderly people, that live alone. The recognition of postures is a key step in the global process of activity recognition. In recent years, many approaches have been proposed addressing this problem. Posture recognition usually relies on : (1) mono-camera systems, or (2) multi-camera systems.

In the first category, (Kellokumpu et Heikkilä, 2005) proposed a mono-camera system for human activity recognition by modeling activities as a continuous sequence of

discrete postures using Hidden Markov Model. (Girondel et al., 2005) developed a system that can automatically recognize different static human body postures in video sequences. The recognition is based on feature fusion using the belief theory. (Nadia et al., 2008) developed a monitoring system to recognize a set of activities of elderly at home by using ten 3D key human postures. The recognition process is based on comparing the detected person's silhouette with each 2D projection of the 3D key postures. The most similar key posture represents the current posture of the observed person. The authors in (Onishi et al., 2008) presented a method to estimate the 3D human postures expressed by a multi-joint model, using the HOG features from a monocular image. (Juang et Chang, 2007) proposed a human body postures recognition approach based on a neural fuzzy network applied to detect emergencies that are caused by accidental falls. In addition, (Pham et al., 2007), used the technology of thermal imaging to develop a video-surveillance system for posture analysis in a crowd.

The mono-camera systems mentioned above can be adapted for different environments and have a low time complexity. However, these methods are dependent of the viewpoint and very sensitive to occlusion and ambiguities. Furthermore, different 2D body postures may be very similar under perspective projection. This constraint limits the use of 3D techniques in applications utilizing monocular system.

In the second category, multi-camera systems are used to deal with problems of mono-camera systems. (Wu et Aghajan, 2007a) developed a system for gesture recognition using an opportunistic fusion smart camera network. Besides, (Pellegrini et Iocchi, 2007) presented a system for posture tracking and classification based on a stereo-vision sensor. Both methods are based on matching 3D data with a 3D human body model. (Cohen et Li, 2003) used a 3D visual-hull constructed from a set of silhouettes for inferring the 3D body postures. They introduced an appearance-based, view-independent, 3D shape descriptor for classifying and identifying human posture using a support vector machine. (Chu et Cohen, 2005) characterized an arbitrary human posture as a linear combination of primary and secondary atoms. These atoms are represented through their shape descriptors extracted from the 3D visual-hull of the human body posture.

Correct classification rate in posture recognition are generally better when multiple cameras from different viewpoints are used, yet most of conventional methods have used one camera, because of the reduced cost and fewer technical difficulties including (i) easier installation in real environments (ii) less camera and no need to synchronize cameras, (iii) avoid network problems (loss of data, limited bandwidth, etc), (iv) reduction of computation costs. Hence, to deal with these limits and still benefit from the accuracy advantages of multiple camera approach and the simplicity of the single camera, one of options is to design a "multi-view" system using a single camera and cast shadows obtained through illumination by few lights.

In fact, a light source can be considered as a special case of a camera, which generates an image of a cast shadow. This image presents the projection of the person's body posture.

Moreover, vision applications often integrate cast shadows into their models, either by treating them as noise to be detected and ignored (Guo et al., 2013a), exploiting them for camera calibration (Cao et Shah, 2005a, Junejo et Foroosh, 2008), incorporating them into larger image formation models (Ackermann et al., 2012a) or exploiting their inherent structure to recover shape from a single view (Abrams et al., 2013a, Gouiaa et Meunier, 2014a). In addition, Yumi et al. proposed a list of paper (Iwashita et al., 2010a, 2012a, 2013, 2012b, Shinzaki et al., 2015) in which they introduced shadow biometrics methods for person identification in the context of security in controlled spaces. Precisely, cast shadows projected on the ground either by the sun in daytime or lights during the night, and captured by a mono camera are used in combination with body silhouette for gait analysis in order to identify persons.

In this paper, we propose an hybrid system (multi-view with a single camera) for human posture recognition based on the fusion of the body silhouette with cast shadows of the person. In the proposed system, we install 4 infrared lights at the different upper corners of a room, and a camera with an infrared filter in the ceiling to capture images. The lights are turned on and off cyclically using a simple electronic device to get one shadow by frame. The advantage of using infrared lights is that these lights do not disturb

the subject. Cast shadows of the person projected by multiple lights can be considered as body silhouettes captured from different viewpoints. Features representing human postures are extracted from the fusion of each cast shadow and the person's silhouette. A weighted majority vote scheme is used to decide which is the corresponding posture.

This paper is organized as follows : Section 2 introduces the overall system and the proposed algorithm with multiple infrared light sources. Section 3 shows experimental results and analysis obtained by testing the proposed algorithm on a dataset captured in our laboratory. Finally, conclusions and future work are presented in section 4.

3.4 Overall system description

The proposed multi-light human posture recognition, shown in Fig.3.1, consists of a parallel structure where each block is considered as an independent human posture recognition system that processes every frame coming from the corresponding light source. Each step is now described in details below.

3.4.1 Silhouette and cast shadow extraction

By using controlled infrared illumination, we avoid illumination variation problems which cause several issues in classical background subtraction task. Hence, the silhouette and cast shadow extraction are carried out using a straightforward background subtraction technique. First, the current frame acquired by the camera when the scene is illuminated by the light source i is subtracted from the background image corresponding to the light source i $(i = 1,...,4)$. Next, a thresholding operation is applied on the 4 resulting images cyclically to remove their backgrounds, and obtaining 4 images noted (sh_1, sh_2, sh_3, sh_4) (see Fig. 3.2 row 1) containing the person's silhouette and each one of the four cast shadows.

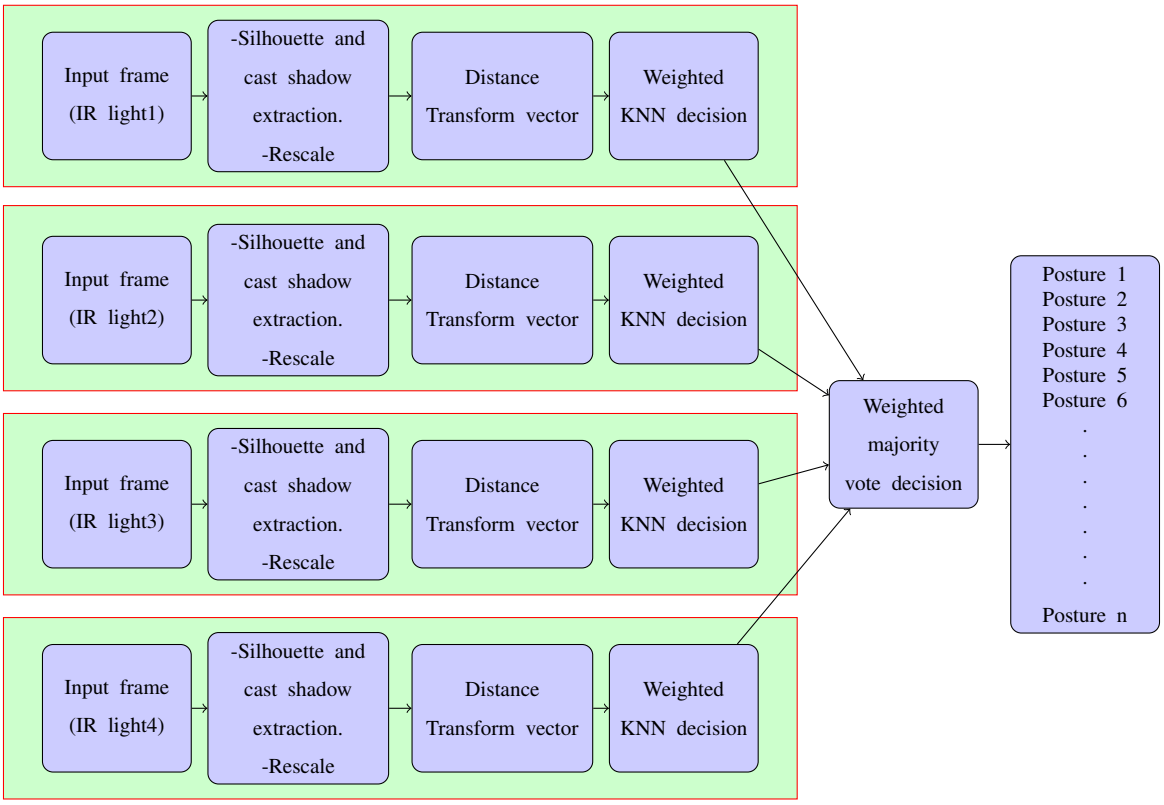


Figure 3.1 – Multi-infrared-light human posture recognition system.

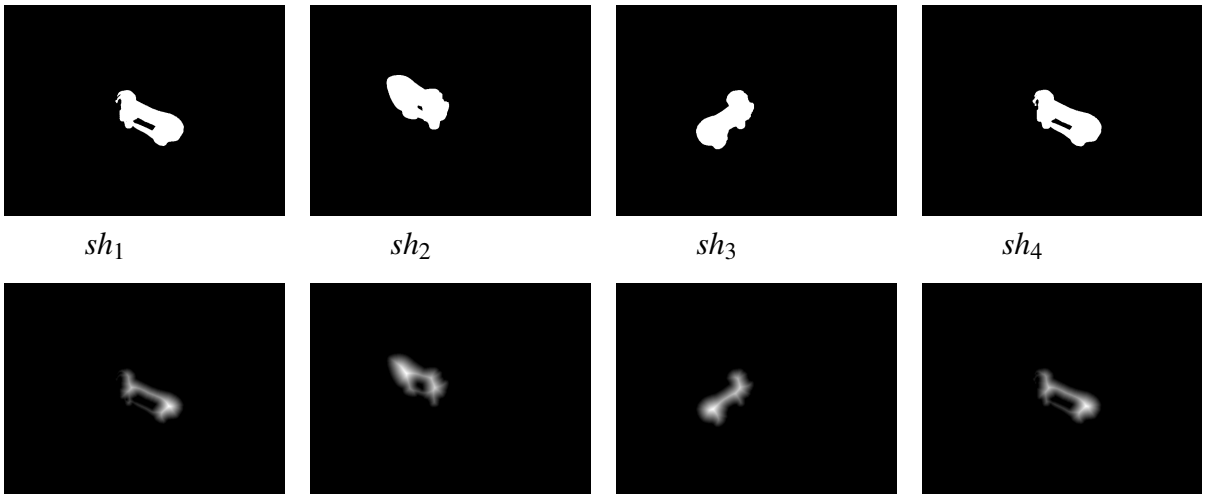


Figure 3.2 – Body silhouette and cast shadows extraction and the corrresponding Euclidean distance transform features.

3.4.2 Feature Extraction

A distance transform of a binary image specifies the distance from each pixel to the nearest feature. The most common feature is a boundary pixel. Several different sorts of distance transform exist, depending upon which distance metric (euclidean distance, chessboard distance, city block distance, etc) is used to determine the distance between pixels. Distance transforms are widely used for comparing binary image and therefore classifying human postures. In our case, we use the approach suggested in (Nater et al., 2010a).

Each image (sh_1, sh_2, sh_3, sh_4) is rescaled to a fixed number of pixels (100×100 in our case) and an Euclidean distance transform is applied. Pixel values are normalized by the maximum distance value, and the rows are concatenated in a vector that defines the fixed length image features $(n = 10000)$, describing the appearance of one person with its cast shadow in the scene (see Fig. 3.2).

3.4.3 Posture classification

To evaluate the discriminatory capabilities of the extracted features in each block, we use the popular and simple (Wu et al., 2008) K-nearest neighbor (KNN), which finds a group of K objects in the training set that are closest to the test object, and assigns a label based on the predominant class in this neighborhood. Majority voting can be expressed as :

$$y^{'} = \arg\max_{v} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i) \tag{3.1}$$

where v is a class label, y_i is the ith neighbor's class, $y^{'}$ is the class of the test object, D_z is the set of the selected nearest neighbors, $\mathbf{x}_i$ is the feature vector of a training object $(\in D_z)$ and $I(.)$ is an indicator function that returns 1 if its argument is true and 0 otherwise.

The straightforward majority vote scheme used for combining class labels can be a problem if the nearest neighbors vary broadly. For this issue, we used a more sophisticated approach(Weighted KNN or WKNN) which weights each object's vote by its distance,

where the weight factor is given by the reciprocal of the squared distance between the feature vectors of the test object x' and its neighbors x_i : $w_i = \frac{1}{d(\mathbf{x}',\mathbf{x}_i)^2}$. So, equation 3.1 is replaced by :

$$y^{'} = \arg\max_{v} \sum_{(\mathbf{x}_i,y_i)\in D_z} w_i \times I(v = y_i) \tag{3.2}$$

We use the χ^2 test static as a similarity metric which gave good results in (Nater et al., 2010b). Therefore :

$$d(\mathbf{x}^{'},\mathbf{x}_i) = \frac{1}{2}\sum_{l=1}^{n} \frac{[\mathbf{x}^{'}(l) - \mathbf{x}_i(l)]^2}{\mathbf{x}^{'}(l) + \mathbf{x}_i(l)} \tag{3.3}$$

where $n = 10000$ is the dimension of each feature vector and $x^{'}(l)$ is the lth component of the feature vector.

As depicted in figure 3.1, for each block (an independent classifier), the weighted K- Nearest Neighbor (WKNN) classifier is used to infer the most likely class. Finally, a weighted majority vote technique is used to decide the corresponding posture based on the outputs of multiple classifiers.

3.5 Experimentals results and analysis

3.5.1 Data set

Videos were recorded using our proposed multi-infrared light source system in a $(4m \times 4m \times 3m)$ room, where 4 infrared light sources are attached in the different upper corners and a camera (Prosilica GC1380) with an infrared band-pass filter and a wide angle lens was installed in the ceiling to capture images. In addition, infrared-lights are automatically turned on and off using an electronic device in order to get one shadow by frame. This system is based on the Relay Shield V2 (rel, 2015), which is an Arduino compatible module with 4 mechanical relays. We connected each light source to one mechanical relay and the Relay Shield V2 was directly controlled by an Arduino Uno microcontroller (ard, 2015) to switch the relays and turn on and off cyclically the different light sources.

Because, the infrared waves can be absorbed by some materials, we used a polyester sheet which has a large reflectivity according to (Zhou et al., 2011a), to cover the ground and reduce the effect of the absorption of the infrared radiation. As is the case when a new computer vision approach is explored, a simple scene with no furniture was used in our experimental setup to better evaluate the potential of the approach.Videos were simply recorded in a room where the shadow were projected on the ground.

In this dataset, we considered some of the basic postures of a person proposed by a medical experts as in (Nadia et al., 2008). Our dataset was composed of the following postures : walk (class1), crouch (class2), pickup (class3), sit on a chair (class4), stand up (class5), take an object (class6), sit down with straight legs (class7), sit down with bent legs (class8), laying down (class9) (see Fig. 3.3). Each posture was performed by 6 volunteers at different places and orientations in the room for a total of 162 postures(=162×4 images). The dataset is available from the author upon request.

We carried out two experiments as follows : (i) posture classification with body silhouette and shadows, (ii) posture classification with body silhouette only. The second test is done to compare our system (one camera with four infrared lights) with the classical mono-camera system. We extracted automatically the body silhouette corresponding to each posture using the AND operator between two images sh_i (see Fig. 3.2) with sources installed at opposite corners, then the distance transform feature is extracted from this binary image and one WKNN classifier is used for posture classification. Moreover, for both tests, we used the leave-one-out cross validation technique (leave-one-example-out and leave-one-actor-out) to estimate the classification error rate. The details of each experiment is described in the following sections.

3.5.2 Leave-one-example-out

In this test, we iteratively considered all instances as a labeled data except one and we tested the classifier on the left-out instance. Finally, the average accuracy was calculated over all postures. Using only the body silhouette, we achieved a best accuracy of 70.3% with K=1. Whereas, the accuracy was outstandingly increased to 94.2% by ad-

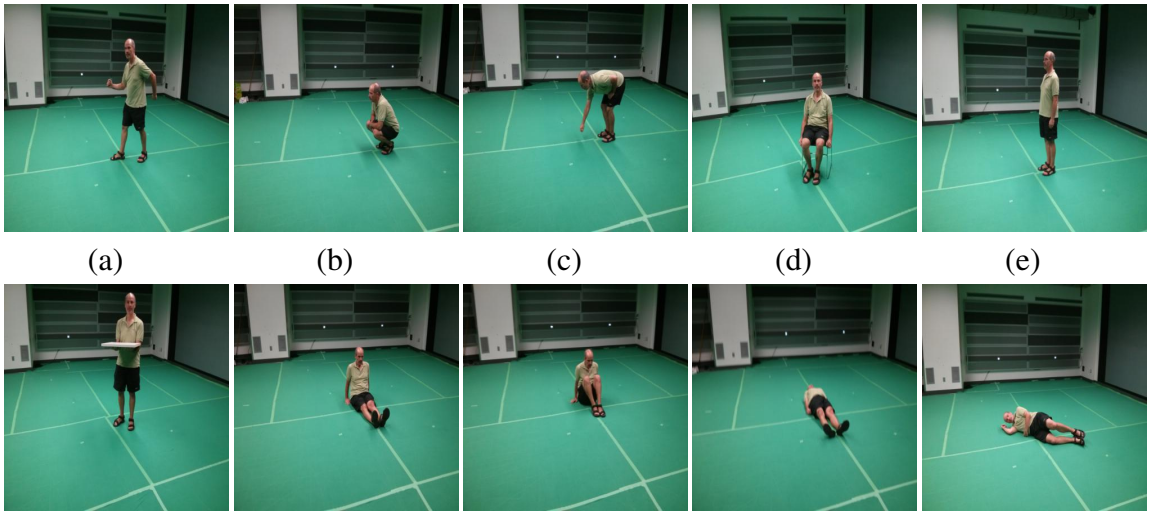
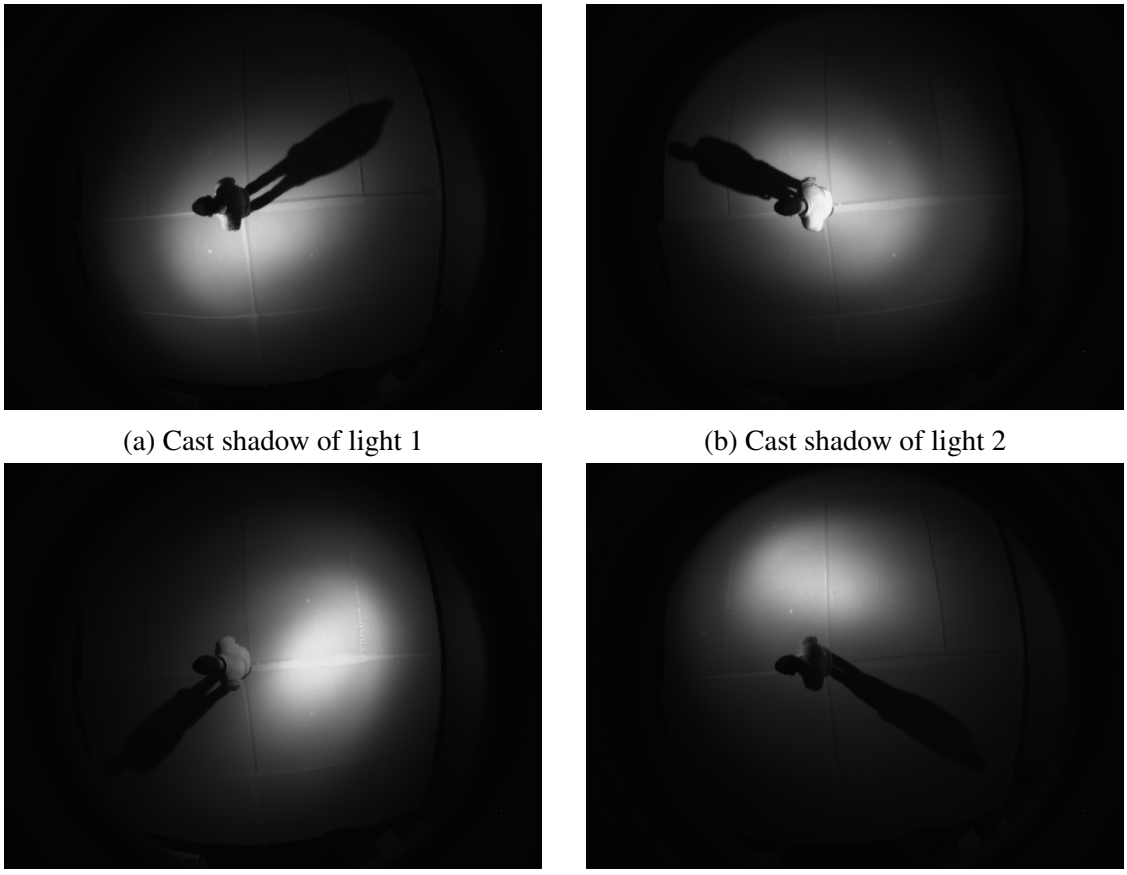


Figure 3.3 – The different posture classes in our dataset : (a) walk, (b) crouch, (c) pick up, (d) sit on chair, (e) stand, (f) take an object, (g) sit down with straight legs, (h) sit down with bent legs, (i, j) lay down

ding shadows and picking K=3 (best results). (see Table 3.I and Table 3.II for confusion matrices results). From this result, it is clear that shadows provide useful information for human posture recognition. In addition, best results using body silhouette and cast shadow were obtained with k=3 $>$ 1, which can be explained because feature points in each class become more compact in the feature space when adding cast shadows. However, the stand up postures were well classified in both cases, which can be explained by these postures are straightforward and can be seen "identically" from different viewpoints. Furthermore, the classification rate of laying postures is sorely decreased, from 94% when using body and cast shadows to 50% when using the body silhouette only, despite the fact that shadow areas are minimal when the person is laying down. We think that this deterioration is due to the high similarity appearance between laying down (class 9) postures and laying down with straight (class 7) legs postures when we eliminate shadows.

Note that our system efficiently classified almost all classes. It is worth noting that many of these silhouettes are very challenging (cast shadows not totally covered by the field of view, cast shadow hidden by the silhouette, silhouette very small when the per-



(a) Cast shadow of light 1

(b) Cast shadow of light 2

(c) Cast shadow of light 3

(d) Cast shadow of light 4

Figure 3.4 – Exemple of real data : stand up posture

son is exactly under the camera, etc). Some examples of these difficulties are illustrated in Fig.3.5. Nevertheless, our approach was able to correctly classify these partially occluded postures using the weighted majority vote scheme based on the output of the 4 classifiers.

3.5.3 Leave-one-actor-out

In this experiment, we iteratively took instances of 5 actors as labeled data and we classified instances of the left-out actor. This task was done to assess the generalization efficiency of our system for "unseen" subjects. This test is also carried out on data with body silhouette and shadows and on data with only body silhouette. As in the previous

(a) (b) (c) (d)

Figure 3.5 – Example of challenging images in our dataset : (a) walk, (b) pick up, (c) stand, (d) take an object.

		Predicted label								
		Walk	Crouch	Pick up	Sit on chair	Stand up	Take an object	Sit down with straight legs	Sit down with bent legs	Laying down
True label	Walk	**14**	1	0	0	2	1	0	0	0
	Crouch	0	**18**	0	0	0	0	0	0	0
	Pick up	0	0	**18**	0	0	0	0	0	0
	Sit on chair	0	0	0	**18**	0	0	0	0	0
	Stand up	0	0	0	0	**18**	0	0	0	0
	Take an object	0	1	0	0	3	**14**	0	0	0
	Sit down with straight legs	0	0	0	0	0	0	**18**	0	0
	Sit down with bent legs	0	0	0	0	0	0	0	**18**	0
	Laying down	0	1	0	0	0	0	0	0	**17**

Tableau 3.I – Confusion matrix of leave-one-example-out test using body silhouette and shadows **(accuracy=94.2% with K=3)**.

test, the best result was achieved when using body silhouette and shadows with an accuracy of 93.1% (K=3). These results are summarized in Table 3.III and Table 3.IV. The efficiency of shadows for pose recognition were also confirmed in this test.

We observe only a slight decrease of accuracy, which confirms that our approach is independent of the individual characteristics of each actor. However, we remark in both tests that the accuracy of class1 (walk) was not very effective, the best result being 66.5%

		Predicted label								
		Walk	Crouch	Pick up	Sit on chair	Stand up	Take an object	Sit down with straight legs	Sit down with bent legs	Laying down
True label	Walk	**9**	2	2	2	3	0	0	0	0
	Crouch	0	**14**	1	1	2	0	0	0	0
	Pick up	0	3	**10**	1	0	0	0	3	1
	Sit on chair	0	0	1	**16**	1	0	0	0	0
	Stand up	0	0	0	0	**18**	0	0	0	0
	Take an object	1	1	0	2	2	**12**	0	0	0
	Sit down with straight legs	0	0	1	0	0	0	**13**	4	0
	Sit down with bent legs	0	1	1	2	0	0	1	**13**	0
	Laying down	0	0	0	1	1	0	6	1	**9**

Tableau 3.II – Confusion matrix of leave-one-example-out test using body silhouette only **(accuracy=70.3% with K=1)**

in test 2 (see Table 3.III). This degradation is mostly due to widely different walk poses of the actors producing a scattered distribution of feature vectors and also because of similarities between "walk", "stand" and "take an object" poses in some cases.

All above tests were also done using the classical KNN classifier and gave about a 1 % lower accuracy than results obtained with WKNN.

Class	Average accuracy (%)
Walk	66.5
Crouch	100
Pickup	100
Sit on chair	100
Stand up	100
Take an object	77.33
Sit down with straight legs	100
Sit down with bent legs	100
Laying down	94.33
Average = 93.1 with K=3	

Tableau 3.III – Leave-one-actor-out results using body silhouette and shadows.

Class	Average accuracy (%)
Walk	44
Crouch	77
Pickup	55
Sit on chair	88
Stand up	100
Take an object	66
Sit down with straight legs	61
Sit down with bent legs	83
Laying down	33
Average=67.4 with K=1	

Tableau 3.IV – Leave-one-actor-out results using body silhouette only.

3.6 Conclusion and future work

We proposed a mono-camera system with multi-infrared lights for human posture recognition in which body silhouette and shadows projected onto the ground are extracted automatically from captured images. We then analyzed posture features extracted from the combination of body silhouette and shadows using the distance transform. Posture classification are done using a weighted majority vote scheme. We carried out experiments using a dataset collected in our laboratory using our proposed system. Moreover, we demonstrated that the proposed approach outperformed the conventional mono-camera approach. In addition, we built a new shadow database for human posture recognition. Notice that, thanks to the infrared light sources, our system can work day and night, and does not bother the subject since IR light is invisible.

A simple scene with no furniture was used in our experimental setup to better evaluate the potential of the approach ; our future work will consider assessing its robustness in a more complex scene with shadows projected on walls, doors, and some furnitures. We also plan to use a fish eye lens (with radial distortion correction) in order to guarantee the full coverage of the scene. In these more complex situations, we expect that the cluster shape of each posture in the feature space will be more complex so that WKNN will not be sufficient, and more powerful machine learning approach and larger data set will be necessary to resolve this problem. In addition, a more sophisticated (than distance transform) descriptor could be necessary to improve the classification of some postures. Besides, in this work, the classification was done by a majority vote scheme where each light source was considered as an independent camera, however, an important aspect of future research will be the ideal combination of body silhouette and 2, 3 or 4 shadows in the same image for classification. This would greatly simplify our system since the light sources would not need to be turn on/off cyclically.

CHAPITRE 4

LEARNING CAST SHADOW APPEARANCE FOR HUMAN POSTURE RECOGNITION (ARTICLE)

Ce chapitre présente le manuscrit intitulé "*Human posture recognition by combining silhouette and infrared cast shadows*" soumis au journal *Pattern Recognition Letters* par Rafik Gouiaa et Jean Meunier.

4.1 Avant-propos

Dans notre travail précédent, (Gouiaa et Meunier, 2015b), nous avons montré que l'ombre projetée directement sur le plancher fournit des informations pertinentes pour la reconnaissance de postures humaines. En plus, (Iwashita et al., 2013, 2012b) ont auparavant montré l'importance de la combinaison de l'ombre projetée (par le soleil ou par une lumière artificielle) avec la silhouette du corps dans le contexte de l'identification de personnes à partir de l'analyse de la démarche. Cependant, ces méthodes ont supposé des scènes très simples (sans murs, sans meubles, etc.) où l'ombre se projette directement sur le plancher sans avoir aucune déviation ce qui n'est pas très réaliste.

Dans ce travail, nous installons donc notre système multivues, basé sur une caméra et un certain nombre de lumières infrarouges, dans un environnement plus réaliste (couloir) pour la reconnaissance de la posture. Dans ce cas, chaque ombre peut se projeter sur différentes surfaces (murs, plancher) en générant des projections complexes de corps qui représentent différentes formes dans la même classe de posture. En outre, ces images paraissent difficiles à décrire par de simples descripteurs de formes qui nécessitent certaines invariances de formes dans la même classe. Nous proposons donc d'utiliser un réseau de neurones convolutionnel (*CNN*), qui est connu pour sa capacité d'apprentissage d'une meilleure représentation des données à partir d'une large base de données. En l'absence de suffisamment de données réelles, nous avons simulé une grande base de données synthétique pour faire l'apprentissage de notre modèle *CNN* en utilisant une

technique permettant de normaliser les données synthétiques et réelles. Notre modèle a été évalué sur des vidéos réelles et produit de meilleurs résultats que la méthode développée au deuxième chapitre.

Les notations utilisées dans cet article sont liées à l'article et n'ont pas de lien avec le reste de la thèse.

4.2 Abstract

This paper presents a system for human posture recognition using a camera and two infrared light sources. It uses as input the combination of the body silhouette and its (invisible to the eye) cast shadows. Conventional video-surveillance methods based on a single camera can fail to infer the correct posture since different postures can look similar under perspective projection. Fortunately, cast body shadows, generated by infrared lights, offer additional posture information that cannot be directly captured by a single camera. Each shadow can be projected on different surfaces (e.g. floor, walls and furniture) generating complex body projections that represent various shapes within the same posture class. These shadow images are very challenging and difficult to describe with traditional handcrafted features that need to be somewhat invariant to these within-class changes. However, a deep convolution neural network (CNN) is able to learn a better data representation from a large-scale dataset. In the absence of a big real dataset, we propose to use synthetic data for training the CNN classifier. Learning from synthetic data is a challenging task due to the gap between synthetic and real feature distributions. Thus, we propose a normalization technique to bridge the gap and help the classifier to better generalize with real data. We evaluated the proposed system on a new real dataset captured in our laboratory and a simulated dataset generated with computer graphics tools. Experimental results validated the efficiency of the CNN model against other conventional methods. Furthermore, the combination of cast shadows and body silhouette had better performance than using only the body silhouette as expected.

Keywords : Cast shadows, human postures classification, convolution neural network, video surveillance, transfer learning, infrared

4.3 Introduction

We introduce a system for human posture recognition, in an indoor setting, which uses information from the person's silhouette and cast shadows. Such a system can be installed in an uncluttered environment such as a corridor for monitoring relevant activities. Posture recognition is motivated by a wide range of promising applications including human machine interaction, video annotation, elderly home care, home and public area security, etc. In general, posture recognition approaches can be divided in two main categories : (i) posture recognition using one camera, (ii) posture recognition using multiple cameras. Traditional approaches that use input information from one camera typically works as follows : after background subtraction and several preprocessing operations, one or many descriptors such as height of a bounding box (Qian et al., 2010), spherical harmonics (Razzaghi et al., 2013), distance transform (Nater et al., 2010b) are applied to extract features from the minimum bounding box of the resulting silhouette. These features are fed to a classifier in order to predict the class of the posture. While these monocular approaches were shown to be applicable for different environments and have a low computational complexity, their performance is view-dependent and very sensitive to self-occlusion and ambiguities. Moreover, the recovered 2D posture is often not accurate due to perspective projection, where different postures can be similar.

To deal with the limits of mono-camera systems, several approaches were proposed with multiple cameras. They generally combine visual data from different cameras (Srivastava et al., 2014a), or use 3D reconstruction (Chu et Cohen, 2005, Wu et Aghajan, 2007b), for feature extraction and prediction of the posture class. In multi-camera systems, the rich visual data provided by different views allows a better interpretation of posture. Furthermore, higher accuracy in posture recognition and robustness to self-occlusion are generally achieved when multiple cameras are used. Nevertheless, most conventional methods have relied on a single camera to avoid technical issues since they (i) are easy-to-install in a real environment, (ii) do not require camera synchronization, (iii) avoid network problems (limited bandwidth, loss of data) and (iv) use simpler and faster algorithms. One solution to deal with these problems and still benefit from the ad-

vantages of a multi-camera system is to design a “multi-view system" using one camera and cast shadows obtained by a few infrared light sources. In such systems, the infrared lights are installed around the scene to cast the person’s shadows while a single camera is used to capture the person’s silhouette with the cast shadows. The shadows projected by the infrared lights can be considered as a body silhouettes captured from different viewpoints. Hence, the system offers multiple silhouettes captured from multiple viewpoints through only one camera.

Such a system was proposed firstly by (Iwashita et al., 2010a, b, 2012b), (Shinzaki et al., 2015) for person identification in public areas (e.g. airport) for security application. The method combines the person’s silhouette with cast shadows projected on the ground by sunlight or artificial lights in order to analyze gait and identify persons. We proposed a similar system consisting of four light sources and one camera for posture recognition through two different strategies. First (Gouiaa et Meunier, 2014a, 2015a), we predicted the body posture class using a 3D Visual Hull recovered from the person’s silhouette and cast shadows. Second (Gouiaa et Meunier, 2015b), we extracted distance transform features directly from the combination of the person’s silhouette and cast shadows for body posture recognition.

All methods reviewed above achieve high classification accuracy despite the simple features and classifiers used for describing and classifying postures. This success can be explained because they assumed a very simple scene (floor without walls) where cast shadows were projected directly on the ground, and lights were installed at a very high height to get an adequate cast shadows (close to the silhouette size). Otherwise the shadows could be very long with poor contrast. In a more realistic scene, cast shadows will typically be partially projected on both floor and walls with various shapes for the same posture. Furthermore, lights should be placed at a realistic height for an in-home scene. The application of these real-word constraints can greatly affect the accuracy of any method.

To cope with these limits, this paper presents an extension of our previous work (Gouiaa et Meunier, 2015b). First, we consider a more realistic scene, a corridor in a

home or other environment, where cast shadows might be partially projected on walls. Second, to limit the height of infrared lights to a realistic value and still get adequate cast shadow dimensions, we simply embedded a large acrylic mirror in the ceiling and placed the infrared lights below and pointing to it. This way, the mirror created two virtual lights which illuminated the scene from a higher height and generated sharp shadows with a reasonable size (see Fig. 4.1).

In a corridor, each shadow can be projected on different surfaces (i.e. floor, walls) generating complex body projections that represent various shapes within the same posture class. These shadow images are very challenging and difficult to describe with traditional handcrafted features such as distance transform (Gouiaa et Meunier, 2015b, Nater et al., 2010b) and Hu moments (Bobick et Davis, 2001) because of the non-affine relationship between them. A potential solution is to use a deep convolution neural network (CNN) (Bengio, 2009, Ciregan et al., 2012, Jia et al., 2014) which is especially well designed for feature learning and images classification. CNN is a large neural network that requires a huge amount of labeled training data to be built and to get accurate results. However, in real-world scenarios, one always has to struggle to collect such huge dataset.

To cope with the lack of training data, different methods have been proposed (Oquab et al., 2014, Yosinski et al., 2014) to transfer the knowledge gained while solving one task and applying it to a different but related task. Nevertheless, for our case such shared knowledge is non trivial because of the absence of other real public dataset related to our problem. To solve the issue from another angle, certain approaches (Zhang et al., 2015a, b) proposed the idea of using synthetic data associated with real data. However, learning a classifier from synthetic data is very challenging due to the shift in feature distributions between real and synthetic data, called *synthetic gap*. To overpass the gap, we propose a normalization technique to make distributions statistically closer and help the classifier to better generalize on real data.

The rest of this paper is organized as follows : section 2 describes the proposed multi-view infrared light system and the method used for posture recognition. Section 3 presents our dataset and experimental results. Finally, we conclude the study in section 4.

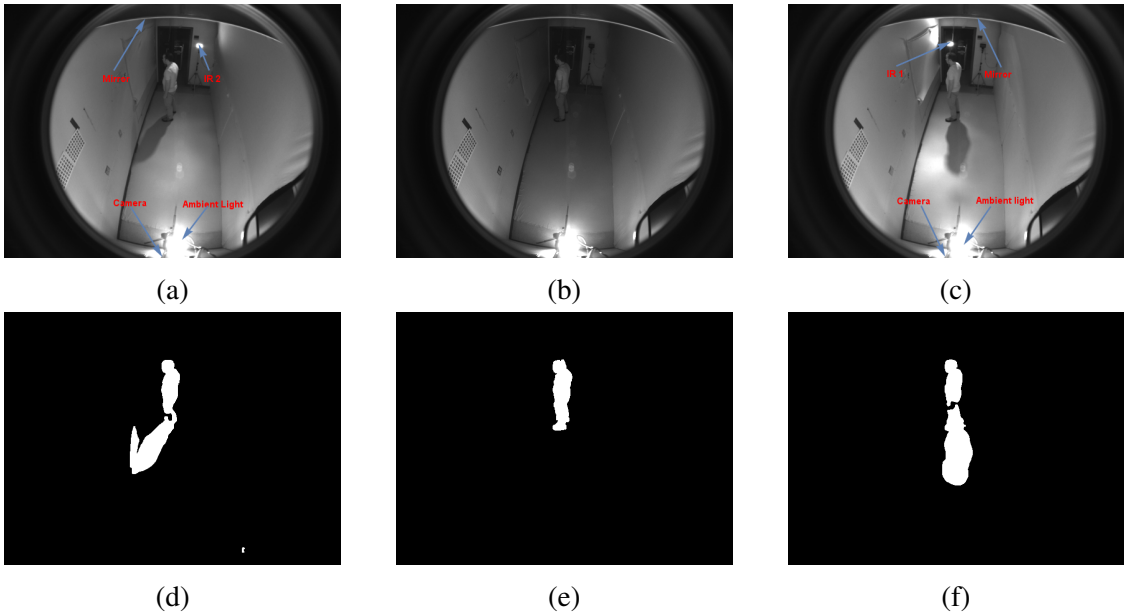
4.4 The proposed approach

4.4.1 Motivation and system setup

We have implemented the system to monitor an individual in a corridor or similar environment. As depicted in Fig. 4.1, the system consists of one camera mounted in the ceiling at one end of the corridor, one large acrylic mirror fixed to the ceiling, and below it, two infrared lights that project light directly towards the mirror consequently creating two virtual lights that illuminate the scene from a height of 1 meter above the ceiling. This enables to obtain sharp cast shadows similar in dimensions to the person's silhouette. To ease processing, infrared lights were automatically turned on and off using an electronic device to get one cast shadow by frame and eliminate shadow superposition. We used a Relay Shield V2 Arduino-compatible which contains 2 mechanical relays. Each light source was connected to one relay and the whole circuit was controlled by an Arduino to switch the relays and turn on cyclically each light in turn. This device enables to get cyclically three different frames : one frame contains the person's silhouette and the shadow generated by light 1 (frame 1), another frame with only the person's silhouette (frame 2), and the last frame with the person's silhouette and shadow generated by light 2 (frame 3) (see Fig. 4.1).

4.4.2 Background subtraction and data normalization

Synthetic data could be an interesting alternative for training classifiers when the amount of real labeled data is not sufficient. However, as mentioned previously, using synthetic data is very tricky due to the gap between the feature distributions of synthetic data versus real data. In our case, the gap can be explained by the fact that images generated under infrared illumination vary a lot with the various types of surface of the scene and textile of human clothing and are difficult to replicate with synthetic data. To deal with the gap problem we decided to normalize real and synthetic images. In any image (synthetic or real) which contains a person's silhouette and two cast shadows, pixels belonging to the person's silhouette were set to 1, while pixels of cast shadows were set



(a) (b) (c)

(d) (e) (f)

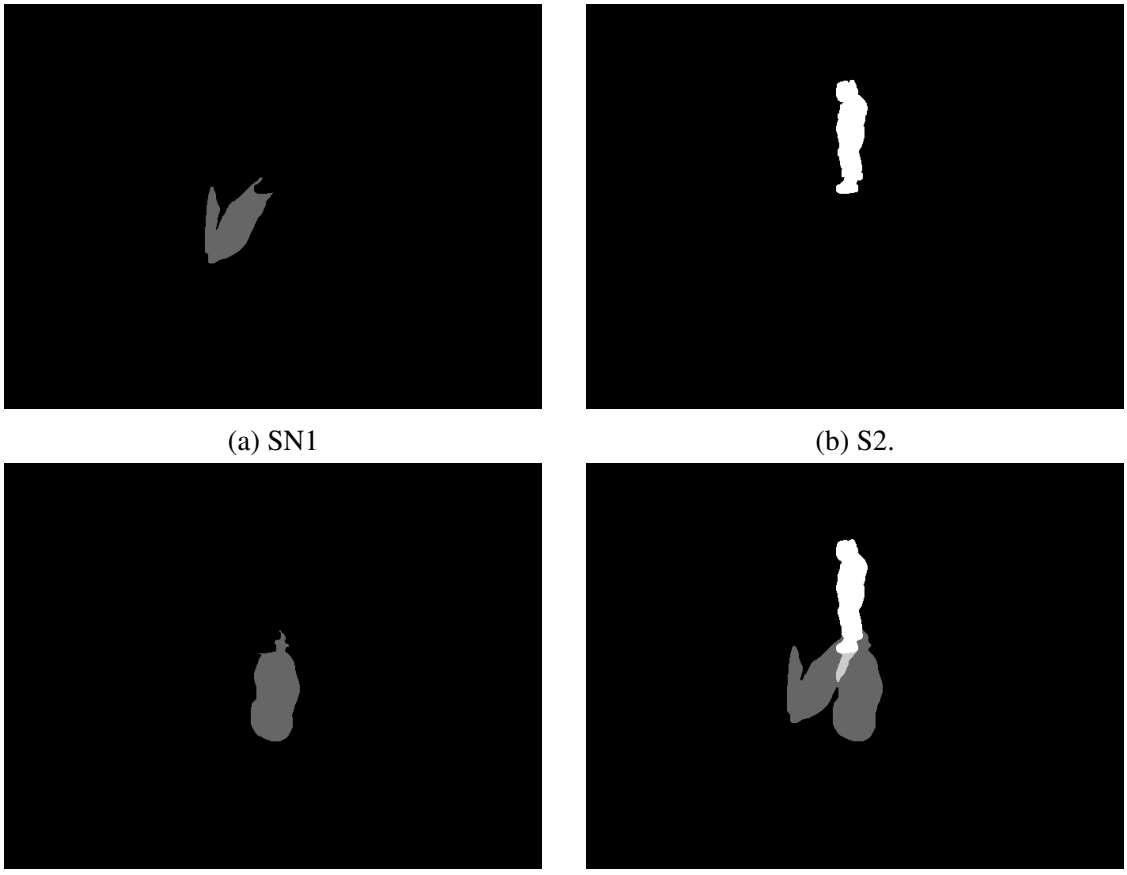
Figure 4.1 – Frames and corresponding binary images of silhouettes generated in one cycle. (a) frame 1, (b) frame 2, (c) frame 3, (d) S1, (e) S2, (f) S3
(see section 4.4.2).

arbitrarily to 0.4 (see Fig. 4.2).

Creating such normalized images is straightforward with synthetic data, while with real data, we used a simple background subtraction method to extract both shadows and the person's silhouette. In figure 4.1, we subtract the binary image S2 respectively from the binary images S1 and S3 in order to obtain shadows from frames 1 and 2. Then, pixel values in both cast shadows were replaced by 0.4 to obtain the normalized shadows (SN1 and SN3). Finally, adding SN1, S2 and SN3 gives the final normalized image which was cropped and resized to 64×64 pixels to be used as input to the CNN (See Fig. 4.2).

4.4.3 Convolution Neural Network

A convolutional neural network is composed of one or more convolutional layers usually with a pooling (subsampling) operation and followed by a number of fully connected layers. Let I be the input image of a convolutional layer with dimensions $n \times n$ where $n = 64$ in our case. A convolutional layer is comprised of l different trai-



(a) SN1 (b) S2.

(c) SN3 (d) Input image

Figure 4.2 – Data normalization.

nable kernels (filters) of size $m \times m$ where m is smaller than the dimension of the image. The input image I is convolved with every filter W to generate l feature maps of size $(n-m+1) \times (n-m+1)$. A max-pooling step over a $s \times s$ contiguous region is applied on each feature map to provide a condensed one. After the subsampling layer, an additive bias b_j and a rectifier nonlinearity (ReLU) σ is applied on each feature map. The output of each feature map j is usually written as follows :

$$h^j = \sigma(b_j + W * I) \tag{4.1}$$

where $*$ is called a convolution operation.

Convolutional layers are used as hierarchically stacked feature extractors ; they de-

tect discriminative and useful information with respect to the input image, from basic to higher-level and more abstract features. Pooling layers are commonly used between convolution layers to reduce the number of parameters (and consequently control overfitting) and introduce some robustness to small image translations. The topmost layer is a softmax classifier used for human posture recognition. Higher order features obtained after the last max-pooling layer are eventually flattened into a 1D vector $V = [v_1, v_2, v_3, \ldots, v_N]$ and classified by a softmax classifier. The predicted class is given as follows :

$$P(Y = i|V, \theta, b) = \frac{\exp(\theta_i^\top V + b_i)}{\sum_{j=1}^{C} \exp(\theta_j^\top V + b_j)} \tag{4.2}$$

$$y_{pred} = \arg\max_i P(Y = i|V, \theta, b) \tag{4.3}$$

where C is the number of classes, y_{pred} is the predicted class whose probability is maximal, (θ, b) is the parameters of the softmax classifier. The softmax classifier minimises the cross-entropy cost function which is written as :

$$\mathscr{L}(\mathscr{X}, \mathscr{Y}) = -\frac{1}{k} \sum_{i=1}^{k} \log(P(Y = y^i|V(x^i), \theta, b)) \tag{4.4}$$

where $\mathscr{X} = \left\{x^{(1)}, \ldots, x^{(k)}\right\}$ is the set of input mini-batch samples in the training dataset and $\mathscr{Y} = \left\{y^{(1)}, \ldots, y^{(k)}\right\}$ is the corresponding set of labels for those input examples. $V(x^i)$ is the feature vector corresponding to the training sample (x^i). Forward propagation is achieved using eqs 4.3 and 4.4. In this work, cost function minimization and weights update were performed using the stochastic gradient descent on minibatches of training data samples.

4.4.4 Convolutional Neural Network Architecture and Hyperparameters tuning

A convolution neural network involves many hyperparameter settings resulting in different architecture configurations. We considered different combination by varying the values of hyperparameters.

We varied the number of convolution layers from 1 to 3 and the fully connected layers from 1 to 2. The number of filters l was chosen in relation with the feature map size so that the bottom layers had fewer filters than the top layers. We selected the number of filters l from 4 to 128 with an interval of 10 for all layers. The filter size $m \times m$ ranged from 11×11 to 3×3. The pooling size was set to $s \times s = 2 \times 2$, the learning rate ρ varied from 0.01 to 0.00001 and the number of hidden neurons in the fully connected layers was chosen in this set $\{128, 512, 1024\}$. To avoid overfitting, we consider the dropout (Srivastava et al., 2014b).

4.5 Experimental Results and Analysis

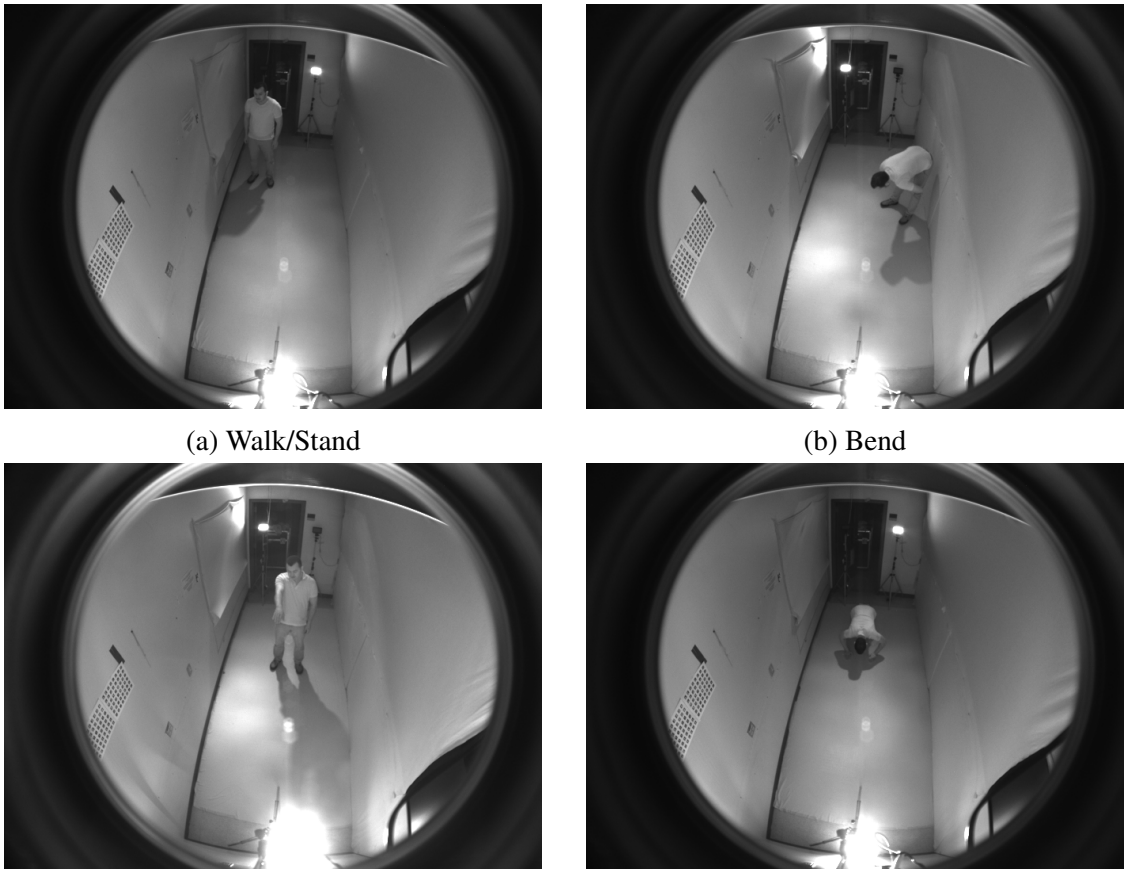
4.5.1 Dataset and Preprocessing

4.5.1.1 Real data set

Experimental videos were recorded using our proposed multi-infrared light source system in a (2m $\times$ 4m $\times$ 2.5m) corridor. We used two infrared lights (niceEshop(TM) 96 LED 12V Night Vision IR Infrared Illuminator) and one monochrome camera (Prosilica-GC1380) with an infrared based-pass filter. We used a large acrylic mirror (2m $\times$ 2m) attached to the ceiling at a height of 2.5m. Both infrared lights were installed below the mirror at 1m from the ceiling and were pointing toward the mirror. So, virtual lights were created at a height of 3.5m from the floor to cast body shadows.

The infrared reflection properties are highly related to the type of material used. To avoid excessive absorption of the infrared radiation by the floor, we also covered the floor using a polyester sheet which has a strong reflectivity according to (Zhou et al., 2011b).

Our dataset consisted of four key human postures (See Fig. 5.4) : Walk/Stand (class 1), bend (class 2), one-hand-stretched (class 3), squat/crouch (class 4). Each posture was performed by 3 volunteers with different body morphologies at different positions and orientations in the corridor for several hundreds of images. The dataset is available from the authors upon request. Fig. 4.1 shows our experimental setup for data acquisition.



(a) Walk/Stand (b) Bend

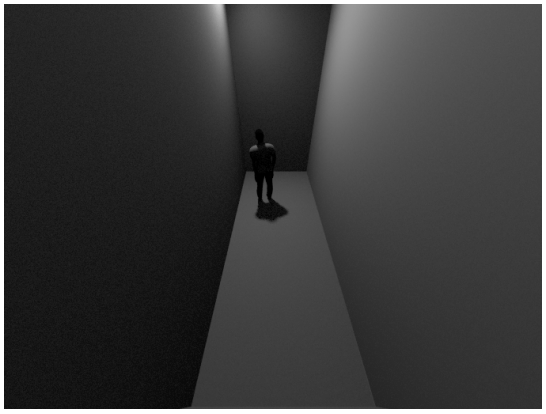
(c) One-hand-stretched (d) Squat/Crouch

Figure 4.3 – The different posture classes in our dataset.

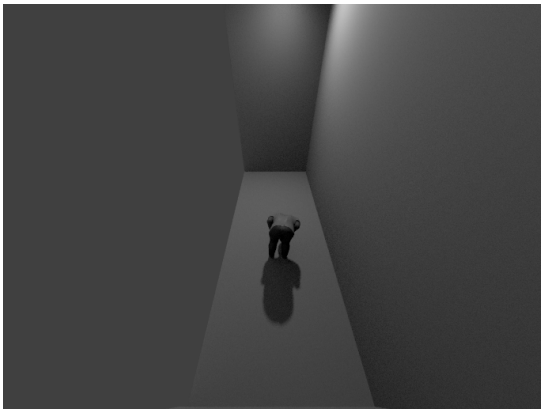
4.5.1.2 Synthetic data set

To generate synthetic data two different computer graphics tools were used : MakeHuman (Team, 2000) and Blender (Foundation, 1995). MakeHuman is a computer graphics software designed for quickly prototyping 3D humanoid characters that can be exported to Blender for 3D modeling and animation in a scene. Three humanoid characters were created to model the volunteer bodies. To imitate the real environment, we considered a corridor with the same dimensions with light sources placed at the same positions as the virtual lights and a camera with the same intrinsic/extrinsic parameters (i.e focal length, sensor dimensions, image size, rotation and translation matrices, etc).

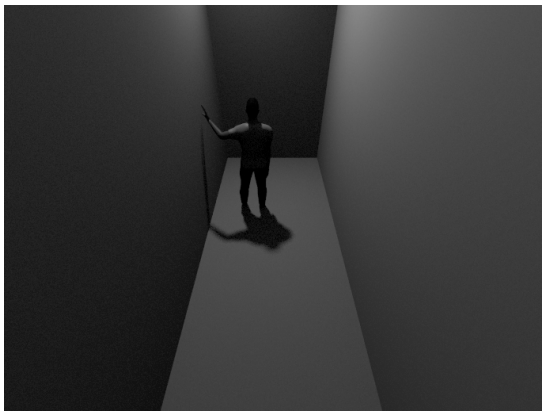
Different Python scripts for Blender were written to randomly move and orientate the



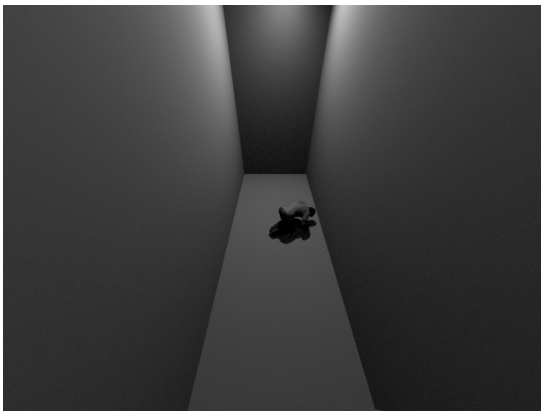
(a) Walk/Stand



(b) Bend



(c) One-hand-stretched



(d) Squat/Crouch

Figure 4.4 – The different posture classes in our synthetic dataset.

humanoid characters in the corridor. For each class, we generated 6000 images (2000 per character). Thus, the size of the synthetic dataset was 24000 different postures used for training the CNN.

4.5.2 Results and discussion

This section presents the results of posture recognition experiments accomplished using the synthetic data set and the real data set collected in our laboratory. Synthetic data were used for training, whereas the real data was used for testing classifiers. We carried out two experiments :

— Posture recognition by combining the person's silhouette and shadows.

Classifier/ Data	CNN	DT+RBF-SVM C=10^3, $\gamma = 0.02$	Hu moments+RBF-SVM C=10^6, $\gamma = 0.0002$	KNN K=9	SVD M=50	MLP
Silhouette +Shadow (Real data) accuracy	**94.58%**	65.14%	76.67%	75.14%	73.51%	72.58 %
Silhouette+Shadow (Synthetic data) accuracy	99%	100%	95%	100%	98%	98%

Tableau 4.I – Posture recognition using person's silhouette and cast shadows.

— Posture recognition using only the silhouette.

The second experiment was performed in order to compare the multi-view light system (one camera with 2 infrared lights) with the traditional mono-camera system to assess the gain obtained with shadow information.

For both experiments, we also compared the CNN model with some conventional methods used for posture recognition based on handcrafted features or raw data : (i) distance transform (**DT**) feature with a RBF-SVM (Nater et al., 2010b), (ii) 7 Hu moments with a RBF-SVM (Hu, 1962), (iii) K-nearest neighbor (KNN) with raw data (Guo et al., 2003), (iv) Multi-layer perceptron (MLP) with raw data (Hippert et al., 2001) and eigenposture method using Singular Value Decomposition (SVD) (Eldèn, 2007).

At first we evaluated a CNN model where the optimal architecture consisted of 4 layers : the first three layers were convolutional and the last one was fully-connected. The number l of filter was fixed respectively to $16, 32, 128$ with size $m \times m = 7 \times 7, 5 \times 5, 3 \times 3$. The max-pooling size was fixed to $s \times s = 2 \times 2$ with 1-pixel stride. The fully connected layer contained 1024 hidden neurons and the learning rate ρ was fixed to 0.001. As a regularization technique, we applied the dropout technique (Srivastava et al., 2014b) on the fully connected layer with a probability of 0.7.

In order to compare the CNN model with some conventional posture recognition methods, we used the 7 Hu moments as features which are calculated from the central moments and are invariant to translation, scale and rotation. We also used the distance transform which specifies the distance of each pixel to the nearest boundary pixel. Each image was centered and rescaled to a fixed number (64×64) of pixel and the Euclidean

distance was used as a metric to determine the distance between pixels. Pixel values were normalized by the maximum distance value, and the rows were concatenated in a vector that defines the fixed length image features. This feature vector is invariant to translation and scaling, but not to rotation. Both feature sets (Hu moments and distance transform) were applied on binary images of silhouette with or without shadows. In both cases, once feature vectors were extracted, we used a linear-SVM for classifying postures. The RBF-SVM classifier was trained using the synthetic data. Hyperparameters γ and C were tuned using the cross-validation and GridSearch technique implemented on scikit learn (Pedregosa et al., 2011). We also evaluated three methods that use directly raw data. We implemented the force brute K-Nearest Neighbor algorithm that classifies according to the minimal distance to labelled samples ($K = 9$ for optimal result). We also tested a multi-layer neural network with two hidden layers and one softmax output layer. The last technique used “Eigenpostures" (50 eigen postures for optimal results) computed with Singular Value Decomposition. The recognition was accomplished by comparing postures (images) in the lower dimension eigenposture basis (Eldèn, 2007).

Table 4.I shows the results of posture recognition by combining the person’s silhouette with cast shadows. We can remark that the accuracy was always higher with synthetic data (vs. real data) whatever the method tested. This can be explained because the synthetic test data were very similar to the training data and less noisy. In real-world scenario, the CNN model outperformed conventional methods and reached a high accuracy (94.58 %). This was due to the capacity of the convolution neural network to extract knowledge from the synthetic data to better generalize on real data. On the other hand, conventional methods failed to transfer what they gained from synthetic data to perform correctly on real data.

It is worth noting that the CNN model performed well on all classes (Table 4.II). Nevertheless, it missed some postures that were very challenging due to noise, background subtraction errors or partially occluded shadows which led to confusion between classes. Furthermore, despite data normalization, synthetic data may fail to simulate all subtleties of real data.

		Predicted label			
		Walk/Stand	Bend	One-hand-Stretched	Squat/Crouch
True label	Walk/Stand	**151**	8	0	0
	Bend	6	**185**	3	0
	One-hand-Stretched	7	0	**105**	0
	Squat/Crouch	0	14	0	**221**

Tableau 4.II – Confusion matrix : CNN using person's silhouette and shadows

The second experiments was done using only the binary person's silhouette in order to compare our system (one camera + two infrared lights) to a classical monocular system and demonstrate the usefulness of shadow information.

We considered a CNN with the same following properties as in the first experiment : number of layers, learning rate ρ, max-pooling size $s \times s$, dropout probability whereas, the number of filters l was respectively fixed to $4, 8, 64$ with size $m \times m = 7 \times 7, 5 \times 5, 3 \times 3$ for optimal results and the number of hidden neurons in the fully connected layer is fixed to 512.

For the other methods (Hu moments+L-SVM, DT+L-SVM, KNN, MLP, SVD), we considered the GridSearch technique as before to tune hyperparameters.

Table 4.IV presents the results of the second experiments. While the accuracy using the synthetic data was still notable, it sorely decreased for all classifiers when the real data were considered. This drop in accuracy can be explained since different postures can look similar under perspective projection depending on the orientation and position of the person leading to some ambiguities as seen in Table 4.III. In particular, class 2 (bend) and class4 (Squat/Crouch). This experiment confirmed that cast shadow information is discriminative and useful for posture classification.

4.6 Conclusion

We presented a hybrid multi-view system that comprises one camera and two infrared lights for human posture recognition. We evaluated our system in a real scene, where the cast shadows were projected on the floor and walls generating complex body projections

		Predicted label			
		Walk/Stand	Bend	One-hand-Stretched	Squat/Crouch
True label	Walk/Stand	**145**	9	0	5
	Bend	22	**141**	1	30
	One-hand-Stretched	4	7	**101**	0
	Squat/Crouch	36	44	4	**151**

Tableau 4.III – Confusion matrix : CNN using person's silhouette alone.

Classifier/ Data	CNN	DT+RBF-SVM C=10^3, $\gamma = 0.02$	Hu moments+RBF-SVM C=10^6, $\gamma = 0.0002$	KNN K=9	SVD M=50	MLP
Silhouette (Real data) accuracy	76.8%	60.2%	42.1%	69.5%	66.6%	65.4%
Silhouette (Synthetic data) accuracy	90.3%	100%	81.0%	99%	99%	92%

Tableau 4.IV – Posture recognition using only silhouette information.

that represent various shapes within the same posture class. These images were difficult to describe with traditional handcrafted features that were not invariant to these within-class changes. A convolution neural network trained on synthetic data was tested to try to solve this issue. We carried out experiments using a real dataset collected in a corridor in our laboratory. We showed that our system performed better than a monocular system as expected, and that a convolution neural network was able to transfer the knowledge from synthetic to real data and outperformed conventional methods. However, our approach missed some challenging postures due to data confusion and background subtraction errors. In our setup, we used two lights to provide enough shadow information without over-complexizing the system. However, this number could range from 1 to a few (e.g. 4 in our previous work (Gouiaa et Meunier, 2015b)). A mirror fixed to the ceiling is a valuable technique to obtain virtual light source when the ceiling height is too low to produce well-defined shadows, this can also be useful to get virtual cameras (instead of lights) for other applications.

In our future work, we will focus on this weakness by considering more accurate normalization data technique and even avoid the background subtraction process given

that CNN model is fundamentally designed for classification of raw images. We will also address the correction of shadow apperance with homography transformation between walls and floor to recover the full shadow's shape on the floor. In this case, handcrafted features could better perform as in our previous work (Gouiaa et Meunier, 2015b). However, this will require an additional calibration steup to compute the homography matrix. Another future improvment is to get rid of the electronic device used to separate cast shadow and silhouette (frames 1, 2, 3). This can be achieved by shadow multiplexing using light sources with different wave length (Cuypers et al., 2009) or with more powerful segmentation algorithms. Finally, to palliate the confusion between different classes, a constrained visual hull 3D reconstrcution could be a promising solution (Gouiaa et Meunier, 2015a).

Acknowledgments

I wish to offer my most heartfelt thanks to my friend Amjad Almahairi for his support and helps for accomplishing this work.

CHAPITRE 5

HUMAN POSTURE CLASSIFICATION BASED ON 3D BODY SHAPE RECOVERED USING SILHOUETTE AND INFRARED CAST SHADOWS (ARTICLE)

Ce chapitre présente le manuscrit intitulé "*Human posture classification based on 3D body shape recovered using silhouette and infrared cast shadows*" publié dans la conférence *Image processing theory, Tools and Applications* (IPTA 2015) par Rafik Gouiaa et Jean Meunier.

5.1 Avant-propos

Les méthodes précédentes se limitent à des informations 2D (silhouette, ombre, contour, etc.) pour classifier des postures ou analyser des comportements. Cependant, dans ce contexte, des informations 3D peuvent être plus discriminantes puisqu'elles nous permettent de mieux localiser par exemple une personne dans une scène ou de reconstruire entièrement l'enveloppe convexe de son corps.

Dans le contexte de la reconnaissance de postures humaines en utilisant la combinaison de silhouette et de l'ombre projetée par une source infrarouge, nous proposons cette fois-ci une méthode pour reconstruire la forme 3D du corps de la personne. Celle-ci est inspirée de l'approche *Shape From Silhouette* qui nous permet d'estimer l'enveloppe visuelle (*Visual Hull* ou *VH*) d'un objet entourée et filmée par plusieurs caméras. Nous considérons une scène simple où l'ombre se projette directement sur le plancher. Étant donné le *VH* déjà reconstruit par la technique de voxelization, nous utilisons un descripteur de forme 3D capable de générer un vecteur des caractéristiques utiles pour la classification de postures. La classification se fait par un classifieur de type K-plus proches voisins pondérés.

Les notations utilisées dans cet article sont liées à l'article et n'ont pas de lien avec

le reste de la thèse.

5.2 Abstract

We introduce a new mono-camera system with multi-infrared lights for human posture recognition, which is based on the 3D body shape recovered from the body silhouette and cast shadows. We propose a new voxelization method inspired from the *Shape From Silhouettes (SFS)* approach for *Visual Hull (VH)* reconstruction. Our setup consists of 4 infrared lights installed in the different upper corners of a room or a corridor, and a camera with an infrared transmitting filter placed in the ceiling. Light sources are turned on and off, cyclically, using a simple electronic system to get one shadow by frame. To illustrate the feasibility of our approach, we used a simple scene, in which shadows are projected directly on the ground. The *VH*s were reconstructed using the voxelization technique. Features were extracted directly from the *VH* using an invariant shape descriptor. A Weighted-KNN classifier were used for classification. Promising results were provided with a classification accuracy of 91.6%.

Keywords : Cast Shadows, Infrared illumination, Posture recognition, Visual Hull (VH), Shape From Silhouettes (SFS), Video surveillance.

5.3 Introduction

This paper proposes a new system to recover and recognize 3D human postures using a new voxelization technique. It is inspired from the Shape From Silhouettes (SFS) approach but using body shadows projected by multiple infrared lights. Typically this system could be used for monitoring people activity in a room or a corridor for instance an elderly that lives alone at home to assess his activities of daily living, detect/prevent falls etc.

The recognition of human posture is one important step of the global process of analyzing human activity and behavior. For this reason, human posture recognition has received a significant amount of attention in the computer vision research community

in the few past decades. This has been motivated by the ambitious goal of achieving automated and real time systems in different areas such as video surveillance, human computer interaction, etc. Several methods have been proposed for the estimation and analysis of the full-body structure. Some approaches have been proposed to infer 3D human body posture from a monocular camera using a human body model (DiFranco et al., 2001, Fossati et al., 2010) , temporal templates (Agarwal et Triggs, 2006a, Nadia et al., 2008) or learning-based method (Agarwal et Triggs, 2006b). However, these techniques are challenging since only the 2D projection of arbitrary poses are acquired with occlusions and ambiguities. Multi-camera systems are required to avoid these problems and recover a more accurate 3D human posture.

Several multi-camera systems have been proposed for estimating 3D human postures. Among these techniques, there are *Shape From Silhouettes* (Chu et Cohen, 2005, Cohen et Li, 2003, Pierobon et al., 2005), stereo vision approaches (Pellegrini et Iocchi, 2007, Ziegler et al., 2006), 3D body scanners (Werghi et Xiao, 2002). The postures are then characterized using shape descriptors or by characterizing body joint configurations. While the use of multiple cameras provides accurate measurements and can avoid occlusion and ambiguity problems, it has some limits : (i) high cost computations (ii) cameras must be synchronized and calibrated (iii) installation in real environments is complicated. Therefore, to avoid these limits and still benefit from the accuracy advantages of multi-cameras approach and the simplicity of the monocular system, one solution is to design a "multi-view" system using a single camera and cast shadows generated with multiple light sources.

Indeed, a light source can be seen as a special case of a camera, which generates an image of a person as a cast shadow. This image reveals the projection of the current human posture.

In addition, vision applications often incorporate cast shadows into theirs models, either by treating them as a noise to be detected and ignored (Guo et al., 2013b), exploiting them as cues for camera calibration (Cao et Shah, 2005b) and incorporating them into larger image formation models (Ackermann et al., 2012b) or exploiting their inherent

structure to recover shape from a single view (Abrams et al., 2013b, Gouiaa et Meunier, 2014b). Besides, Yumi et al. published a list of paper (Iwashita et al., 2010c, 2013, 2012b, b, Shinzaki et al., 2015) in which they proposed shadow biometrics methods for person identification. More specifically, cast shadows projected on the ground by either the sun in daytime or lights during the night, and acquired by a single camera are used in combination with body silhouette for gait analysis in order to identify persons in the context of security in controlled spaces.

In our previous work (Gouiaa et Meunier, 2014b), we established a new hybrid system (single camera and a few light sources) for human posture recognition based on 3D shape reconstructed by a new voxelization method. In this paper, we refine our system by using only 4 infrared light sources placed in the upper corner of an indoor corridor or a simple room to project shadows directly on the ground, and a camera with an infrared band-pass filter and a wide angle lens installed in the ceiling to capture images. In addition, a simple electronic device is used to turn on and off each light in turn to get one cast shadow by frame. We also validate our method on real data for 3D human posture classification.

The rest of the paper is organized as follows : Section 2 describes the overall system. Section 3 shows experimental results and analysis obtained by testing our algorithm on a dataset captured in our laboratory. Finally, conclusions and future work are presented in section 4.

5.4 Overall system description

Our proposed multi-light human posture recognition system consists of the following steps :

5.4.1 Person's silhouette and cast shadow extraction

First, because images are captured by a wide angle camera, which causes some radial distortions, we use a Matlab toolbox (Bouguet., 2008) to undistort the images and conse-

quently improve the quality of a person's silhouette and associated cast shadows. Then, we need to extract the person's silhouette and the infrared cast shadows generated by the 4 infrared lights. For this purpose, we use a very basic background subtraction technique with thresholding which gives good results since the illumination is well controlled. Given 4 consecutive frames, corresponding to the 4 infrared lights (turned on/off in turn) we get 4 images noted (sh_1, sh_2, sh_3, sh_4) each one showing the person's silhouette and one cast shadow. Then, to simplify the segmentation of the person's silhouette (without shadows), we simply combine these images with a logical AND, the shadows are then easily extracted by removing this silhouette area from the 4 images. Morphological operations are applied to denoise a person's body silhouette and cast shadows. Fig. 5.1 summarizes these principal operations where the cast shadows are colored in red, green, blue and yellow and the person's silhouette is shown in white.

5.4.2 3D visual hull reconstruction

Cast shadows, which are projections of a person's body by multiple lights, can be considered as a body silhouette captured from different viewpoints. To reconstruct the 3D body shape, the *Visual Hull* (VH) is reconstructed in the same way as in the *Shape From Silhouettes* method. The most commonly used reconstruction technique is the voxel-based approach. Briefly, a simple version of this approach is given in the algorithm below.

1. Divide the interest space into a grid of $N \times N \times N$ voxels.
2. Initialize all N^3 voxels as a part of the real object.
3. Check for voxels belonging to the real object.
   (a) For i=1,...,N^3
   (b) For j=1,..., n (cameras)
   (c) Project the voxel v_i on the j^{th} 2D silhouette image plane by the camera projection matrix C_j.
   (d) If the voxel v_i is projected outside the j^{th} silhouette plane then the voxel is dismissed

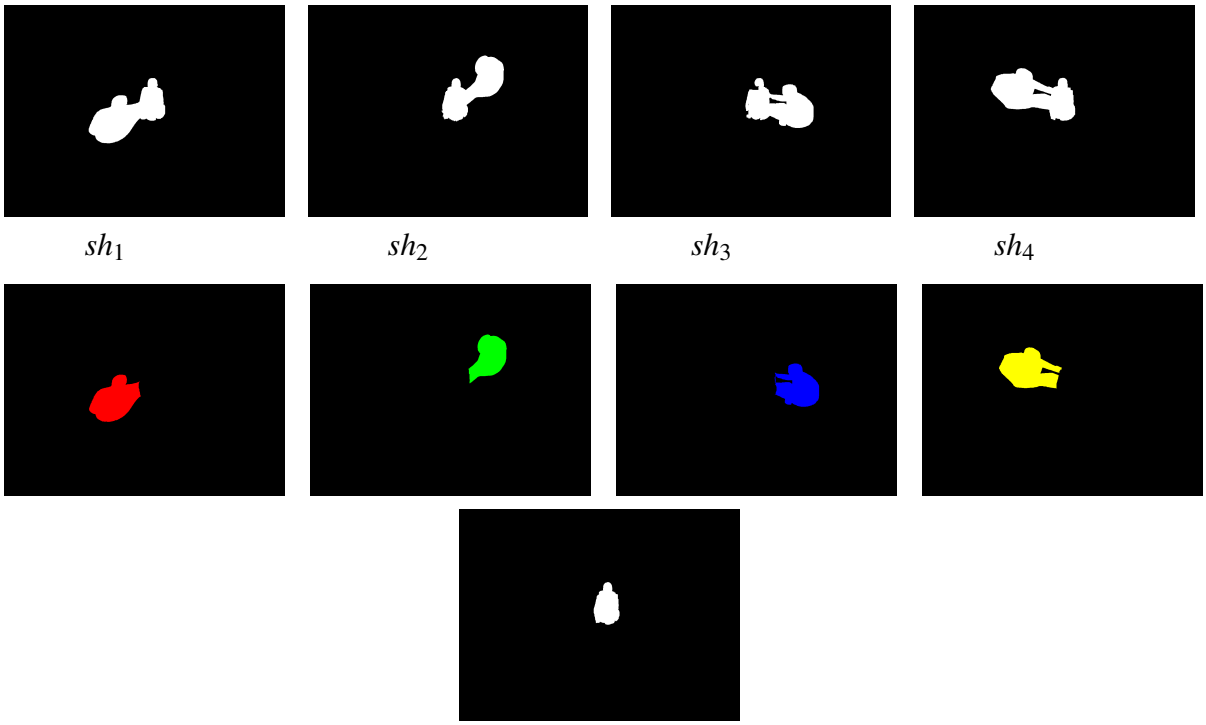


Figure 5.1 – Person's silhouette and cast shadows extraction. The person's body silhouette is extracted by combining images $sh_1, sh_2, sh_, sh_4$ by a logical AND. The shadows are extracted by removing the person's body silhouette from the 4 images.

4. The resulting *Visual Hull* is given by the set of all voxels v_i which are projected inside all 2D silhouettes.

The main idea of *SFS* is to divide the space of interest into a grid of voxels and then iteratively check whether a voxel belongs to the body or not. The *VH* is represented by all voxels projected simultaneously on all 2D silhouettes. However, the step (3-c) is a bit different when we replace a camera by light sources, and is achieved using a simple ray tracing technique. For a light source, the 2D silhouette of step (3-c) is actually the associated cast shadow. A more detailed description follows. Given a light source with a center S_j and a voxel centered in v_i, we consider the 3D line joining S_j and v_i. Its equation is written as :

$$r_{i,j}(t) = S_j + (v_i - S_j)t. \tag{5.1}$$

We extend this line until it reaches the shadow plane where $t = t_0$. The intersection point p is given by :

$$p = S_j + (v_i - S_j)t_0 \tag{5.2}$$

After that, we project p by the single camera projection matrix C on the corresponding j^{th} shadow in the image plane to check whether it falls inside or outside the shadow. Some reconstructed postures are shown in Fig.5.2.

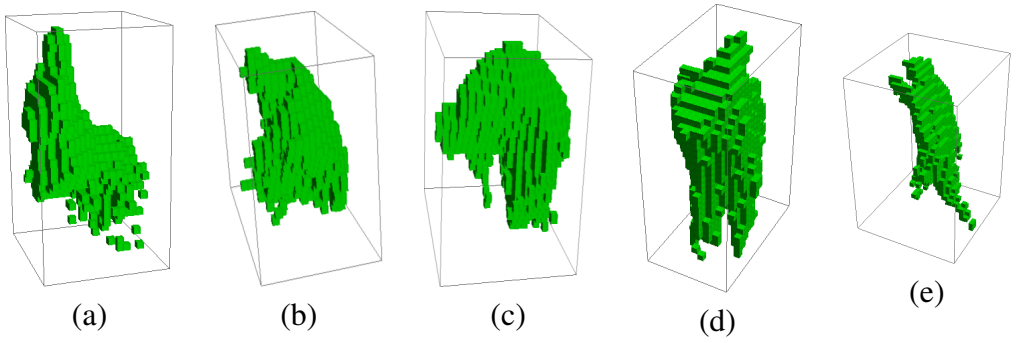


Figure 5.2 – *VH* of some postures in our dataset. (a) sitting on chair, (b) crouch, (c) pick up, (d) stand (e) walk

5.4.3 Shape decriptor

Features representing body postures are extracted directly from the *Visual Hull* using the invariant shape descriptor explained in (Cohen et Li, 2003), and already used in the context of 3D posture human classification in a static environment.
First, we put the *Visual Hull* into a shape reference (e.g. cylinder, sphere). The surface of the defined shape reference is regularly sampled into N *control points*. For each control point p_n, a spherical coordinate system centered on p_n is built, and 3D sectors are defined by creating divisions with dimensions ρ (0 to a maximum value), θ (from 0 to π) and ϕ (from 0 to 2 π). In our case, each polar coordinate is uniformly sampled into ten intervals, obtaining a set of 1000 sectors noted as :
$\{ (\rho_i, \theta_j, \phi_k) : 0 \leq i,j,k \leq 9\}$. For each ($\rho_i$, θ_j, ϕ_k) sector, we count voxels inside and we build the corresponding spherical histogram $h_n(i,j,k)$.
The shape descriptor $H(i,j,k)$ is obtained by summing up the set of histograms $h_n(i,j,k)$ of N control points and normalizing all by the maximum value (for more details see (Cohen et Li, 2003)).

$$H(i,j,k) = \sum_{n=1}^{N} \frac{h_n(i,j,k)}{max_{x,y,z}(\sum_{t=1}^{N} h_t(x,y,z))} \tag{5.3}$$

Using a cylinder as reference shape, we obtain a descriptor ensuring rotation invariance only along its main axis while a sphere provides an invariance rotation descriptor about its center. As suggested in (Cohen et Li, 2003), we used a cylinder as reference shape where its main axis passed through the centroid of the *Visual Hull* vertically oriented and fitted to the body's height. A suitable value of a radius was empirically determined (50 cm in this study). The number N of control points was fixed to 80 (N=80 = 5 longitudinal values $\times$ 16 angular values). An example of the shape descriptor is illustrated in Fig. 5.3

5.4.4 Postures classification

To evaluate the discriminatory capabilities of the extracted features, we use one of the top 10 algorithms in data mining according to (Wu et al., 2008). K-nearest neighbor

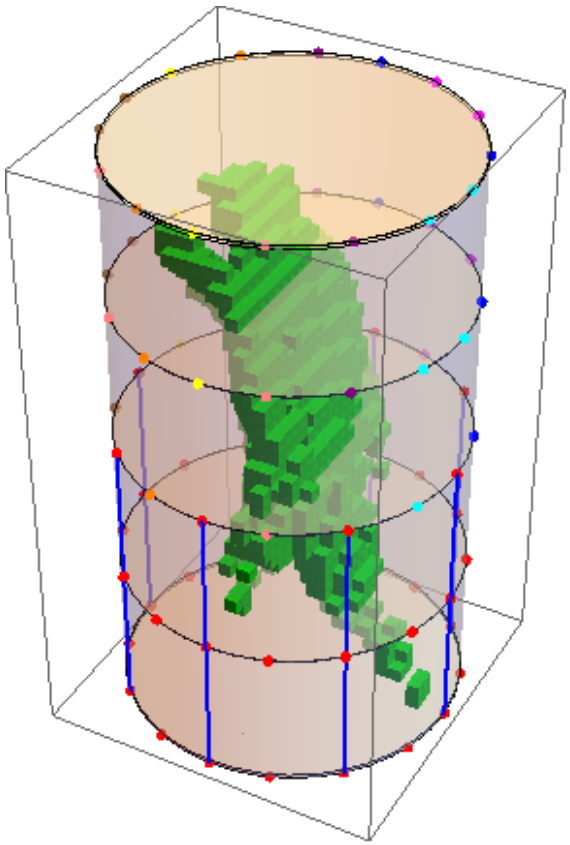


Figure 5.3 – Example of the shape descriptor of a Walk posture.

(KNN) is one of the simplest classifier, which finds a group of k objects in the training set that are more closest to the test object, and assigns a label based on the predominant class in this neighborhood. This approach is based on three key elements : a set of labeled objects, a set of stored records, a similarity metric to compute the distance between samples, and the parameter k, the number of nearest neighbors. KNN method can be summarized by the following algorithm :

- **Input :** D , the set of M training objects, and test object $z = (\mathbf{x}', y')$
- Process :
  - Compute $d(\mathbf{x}', \mathbf{x})$ between z and every object $(\mathbf{x}, y) \in D$.
  - Select $D_z \subseteq D$, the set of k closest training objects to z.

— **Output** :

$$y^{'} = \arg\max_{v} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$$

Given a labeled training set D consisting of a couple data $(\mathbf{x}, y)$ where $\mathbf{x}$ is the feature vector of a training object and y is its class label. To classify the unlabeled object $z = (\mathbf{x}^{'}, y^{'})$, the algorithm calculates the distance between z and all training object in D and the set D_z of the nearest neighbors is therefore selected. Once the nearest neighbors is identified, the label of the majority class in the set of the nearest neighbors is assigned to the class of the test object.

In the majority voting step, v is a class label, y_i is the i**th** neighbors, and $I(.)$ is an indicator function that returns 1 if its argument is true and 0 otherwise.

The straightforward majority vote scheme used for combining class labels can be improved with a more sophisticated approach which weights each object's vote by its distance, where the weight factor is given by the reciprocal of the squared distance : $w_i = \frac{1}{d(\mathbf{x}^{'}, \mathbf{x}_i)^2}$. So, the last step of the above algorithm is replaced by this equation :

$$y^{'} = \arg\max_{v} \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i \times I(v = y_i) \tag{5.4}$$

We used the χ^2 test static as a similarity metric as (Nater et al., 2010a). Therefore :

$$d(\mathbf{x}^{'}, \mathbf{x}_i) = \sum_{l=1}^{L} \frac{[\mathbf{x}^{'}(l) - \mathbf{x}_i(l)]^2}{\mathbf{x}^{'}(l) + \mathbf{x}_i(l)} \tag{5.5}$$

where $L = 1000$ is the dimension of each feature vector and $x^{'}(l)$ is the lth component of the feature vector.

5.5 Experimental results and analysis

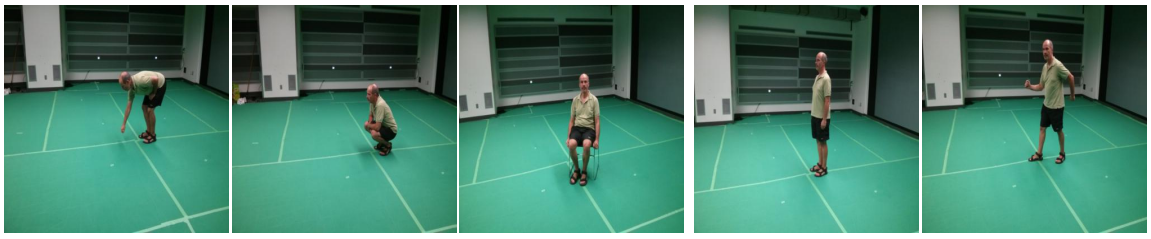
5.5.1 Data set

In this section, we detail our multi-infrared-light system installed in a real environment, and describe some difficulties to acquired realistic videos. Videos were acquired

in a large room $(4m \times 4m \times 3m)$ to ensure that shadows projected on the ground. Four infrared light sources are attached in the different upper corners, while a camera (Prosilica GC1380) with an infrared band-pass filter and wide angle lens was installed in the ceiling to capture images. In addition, we used a simple electronic system to automatically turn on and off each light source in turn and getting one cast shadow by frame. This system was based on the Relay Shield V2(rel, 2015) with 4 mechanical relays and directly controlled by an Arduino Uno microcontroller (ard, 2015) to switch the relays and turn on and off cyclically the different light sources. Because, the infrared waves can be absorbed by some materials, we used a polyester sheet which has a large reflectivity according to (Zhou et al., 2011a), to cover the ground and reduce the effect of the absorption of the infrared radiation.

Our dataset was composed of the following basic postures : pick up (class 1), crouch (class 2), sit on chair (class 3), stand (class 4), walk (class 5) (see Fig. 5.4). Each posture was performed by 3 volunteers at different places and orientations in the room for a total of 60 postures. The dataset is available from the authors upon request.

To test the accuracy and the stability of this approach, we carried out two experiments using the *leave-one-out (leave-one-example-out and leave-one-actor-out)* cross validation procedure.



(a) (b) (c) (d) (e)

Figure 5.4 – The different postures classes in our dataset : (a) pickup, (b) crouch, (c) sit on a chair, (d) stand, (e) walk

5.5.2 Leave-one-example-out

In this test, we iteratively considered all instances as a labeled training data except one and we tested the classifier on the left-out instance. Then, the average accuracy was calculated over all instances. The best result was obtained with k=3 where we achieved an accuracy of 91.6 %. The confusion matrix for the best result is presented in Table.5.I. All postures except Walk were successfully classified. A Walk posture differs from the Stand one in the lower body part only where the legs are further apart. However, in our case, the lower body part in cast shadows can be hidden by the person's silhouette. Therefore, the reconstruction of walk postures was more difficult as shown in Fig. 5.2(e) and often similar to the stand posture.

5.5.3 Leave-one-actor-out

In this experiment, we iteratively took instances of 2 actors as labeled training data and we classified instances of the left-out actor and calculated the average classification rate. This task was done to assess the generalization efficiency of our system for unknown subjects. The best results was achieved by choosing k=3, with an accuracy of 91.6% (see Table 5.2). We note a stability of accuracy, which confirms that our approach is independent of the individual characteristics of actors.

Results with classical KNN (no weight) were poorer by about 3% and 2% accuracy on test 1 and test 2 respectively.

		Predicted label				
		Pickup	Crouch	Sit on chair	Stand	Walk
True label	Pickup	**12**	0	0	0	0
	Crouch	0	**12**	0	0	0
	Sit on chair	0	0	**12**	0	0
	Stand	0	0	0	**11**	1
	Walk	0	0	0	4	**8**

Tableau 5.I – Confusion matrix of the leave-one-example-out test (accuracy =91.6 %)

Class	Accuracy (%)
Class 1	100
Class 2	100
Class 3	100
Class 4	83
Class 5	75
Average =91.6 %	

Tableau 5.II – Average accuracy of leave-one-actor-out test

5.6 Conclusion and future work

We presented a multi-view system consisting of a single camera and multi-infrared lights for human posture recognition. Body silhouette and the different cast shadows projected onto the ground are automatically extracted from captured images. *Visual Hull* of 3D postures were reconstructed using a new voxelization technique inspired from the *SFS* approach, and mainly based on cast shadows. We then analyze the posture feature extracted directly from the *VH* using an invariant shape descriptor. Posture classification is carried out using a KNN classifier. Experiments were done using a data set collected in our laboratory using the proposed system. We showed that the system keeps the advantages of a multi-camera system but is much simpler to implement. Our approach achieved promising results and seems suitable for applications that do not require a very high precision 3D reconstrution. Besides, we created a new shadow database for human posture recognition. In addition, our system can work day and night and does not disturb the subject thanks to the use of invisible infrared light sources.

In this work, we validated the potential of our approach in a simple scene where shadows are projected directly on the ground without furniture ; in our future work, we will consider assessing its robustness in a more complex scene with shadows projected on the walls, doors and some basic planar furnitures. In these more complex situations, the cast shadows will be distorted/broken and we will need to reconstruct the full scene first (get 3D positions of the wall, doors, furniture etc.) to use ray-tracing in our *SFS* algorithm. We will also need to use a fish eye lens (with radial distorsion correction) in order to guarantee the full coverage of the scene. In addition, a more accurate descrip-

tor could be necessary to improve the classification rate of some similar classes such as walk and stand in our previous experiments. The latter problem could also be resolved by using more sophisticated machine learning techniques which would require a larger data set though. In the section 5.4.2, we adapted the basic *SFS* algorithm for *Visual Hull* reconstruction, which supposes that silhouttes and shadows are consistent and complete, and there exists at least one volume that completely explains them. However, cast shadows used in our case are not always consistent/complete due to occlusions (e.g. part of a shadow occluded by the body silhouette). Thus, this algorithm reconstruct only the part of the volume which projects consistently in all cast shadows, leaving the rest unreconstructed. So, in the future, we will extend our voxelization technique to be used with set of inconsistent/incomplete cast shadow. This problem was explored in the case of multiple cameras and a good reference to start with can be found in (Landabaso et al., 2008).

CHAPITRE 6

CONCLUSION

6.1 Synthèse de nos contributions

Dans cette thèse, nous nous sommes intéressés à la reconnaissance de postures humaines par le biais d'un système multivues basé sur une caméra et quelques sources lumineuses infrarouges. Ce système utilise une combinaison de la silhouette de la personne et son ombre projetée par les sources lumineuses afin de reconnaitre une posture dans une séquence vidéo.

Pour traiter le problème de reconnaissance de postures humaines en utilisant un tel système, nous avons proposé deux approches en nous basant sur la dimension de l'espace de travail : une approche 2D et une autre approche 3D.

Nous nous sommes tout d'abord intéressés aux méthodes 2D pour la classification de postures humaines en analysant l'image combinant la silhouette de la personne et son ombre projetée par une source lumineuse (chapitre 3). Pour valider la faisabilité de notre approche, nous avons considéré une simple scène où les ombres se projettent directement sur le sol sans réfraction sur les murs. En plus, nous avons installé les sources lumineuses à une très grande hauteur afin de projeter des ombres de taille raisonnable. Notre système a été évalué sur des vidéos exposant différentes postures humaines et a produit un taux de classification correcte de 94%.

Dans notre deuxième travail sur l'approche 2D (chapitre 4), nous avons proposé d'installer notre système dans un environnement plus réaliste (corridor) où les ombres peuvent se projeter à la fois sur le sol et sur les murs. Les sources lumineuses ont été installées aussi à une hauteur raisonnable pour éclairer un grand miroir fixé au plafond afin de créer des lumières virtuelles éclairant la scène à partir d'une très plus grande distance en générant des ombres de taille proportionnelle à celle de la silhouette de la personne. Nous avons montré que les images capturées dans ce cas sont difficiles à décrire avec

des descripteurs classiques, car, elles révèlent différentes versions non linéairement déformées de la silhouette de la personne. Un réseau de neurones convolutionnel (CNN) a été utilisé pour classifier les postures données en entrée sous la forme d'une image combinant la silhouette de la personne et les deux ombres projetées par deux lumières virtuelles. L'apprentissage du CNN a été fait grâce à une base de données synthétique, alors que son évaluation a été achevée sur des séquences de vidéos réalistes et a atteint un taux de classification correcte de $\sim= 95\%$.

Enfin, l'approche 3D (chapitre 5) a consisté à inférer une posture en utilisant l'enveloppe visuelle construite à partir d'un ensemble de silhouettes (silhouette de la personne+silhouettes d'ombre). Un descripteur 3D basé sur l'apparence de la forme 3D de la silhouette a été utilisé pour classifier et identifier une posture humaine à l'aide d'un classifieur de type k-plus proche voisins (KNN). Notre méthode a atteint un taux de bonne classification de 91.6%.

Malgré la déformation de silhouettes à cause de la projection des ombres sur les murs, l'approche 2D avec l'utilisation de CNN nous permet d'obtenir le meilleur résultat avec un taux de classification correcte de $\sim= 95\%$.

Le montage des expériences en laboratoire a constitué un défi très important dans notre recherche. La recherche et la réservation d'un local approprié et permanent pour installer notre montage a figuré parmi les grands problèmes que nous avons rencontrés durant ce travail. Le choix d'équipement adéquat (caméra, source lumineuse, filtres) a nécessité un grand travail et beaucoup d'expériences. L'installation de la caméra et des sources lumineuses dans le laboratoire n'étaient pas un travail facile à cause de différents problèmes techniques (pas d'attache dans les murs, pas de support pour la caméra...). En plus, le recrutement des volontaires pour réaliser la base de données a demandé beaucoup de planification et un engagement sérieux. Ces différentes tâches aux laboratoires ont présenté une grande partie de notre travail représentant des centaines d'heures de travail.

6.2 Perspectives

À l'issue de ce travail, de nombreuses perspectives s'ouvrent sur les deux approches proposées et sur le système lui-même :

- Dans nos travaux, nous nous sommes limités à la reconnaissance de postures. Cependant, nous pouvons étendre nos méthodes pour la reconnaissance des activités humaines en utilisant une séquence de postures. Ceci peut se faire, par exemple, en modélisant chaque activité par un modèle de Markov caché (HMM).
- L'utilisation de microcontrôleur pour allumer cycliquement les sources lumineuses et avoir une ombre par image peut causer un problème de synchronisation dans le cas de la reconnaissance des activités. Une première idée possible pour éviter l'utilisation du microcontrôleur est d'utiliser des sources lumineuses avec différentes longueurs d'onde (par exemple, NEAR infrarouge vs infrarouge). Dans ce cas, nous pouvons obtenir des ombres de différents contrastes qu'on peut mieux segmenter par un simple seuillage. Une autre solution est d'utiliser des sources lumineuses visibles avec des filtres passe-bandes (filtre pour la couleur rouge, filtre pour la couleur verte...) et de multiplexer les ombres afin d'extraire les silhouettes (silhouettes d'ombre + silhouette de la personne) comme le fait Cuypers et al. (2009).
- L'utilisation du miroir peut résoudre le problème d'installation de sources lumineuses dans un plafond très bas en générant des sources virtuelles. Cependant, l'installation des sources lumineuses (réelles) à une hauteur très basse peut causer une forte illumination de la scène et influencer sur le contraste de l'ombre projetée.
- Dans le cas où l'ombre se projette sur le mur en créant des silhouettes d'ombre déformées, l'utilisation d'un réseau de neurones convolutionnel donne de bons résultats. Cependant, nous pouvons corriger la déformation de l'ombre en utilisant une transformation d'homographies entre les murs et le sol et récupérer la forme exacte de l'ombre projetée. Ceci exige cependant certaines contraintes telles que la connaissance à priori de la géométrie 3D de la scène ainsi que les pa-

ramètres internes et externes de la caméra et la position des sources lumineuses dans l'espace 3D. Une fois qu'on a corrigé la forme, on peut tout simplement utiliser des descripteurs 2D classiques de forme pour achever la reconnaissance de postures.

- La reconstruction 3D de l'enveloppe visuelle d'une posture en utilisant un ensemble des silhouettes d'ombre déformée (à cause des murs) reste aussi faisable avec les contraintes mentionnées auparavant. Une autre approche est d'entrainer un réseau de convolution avec des données synthétiques sous forme des cartes binaires 3D représentant des postures humaines. L'évaluation d'un tel modèle peut se faire par des cartes binaires 3D obtenues à partir des données réelles (silhouettes d'ombre) en utilisant la technique SFS.
- Dans notre deuxième contribution (chapitre 4), nous avons représenté la posture par une seule image qui contient la silhouette de la personne avec deux ombres projetées. On n'est pas obligé, néanmoins, de se limiter à cette combinaison. Nous pouvons proposer, par exemple, de représenter chaque posture par deux images (la silhouette de la personne + une ombre) ou trois images (la silhouette de la personne + image d'ombre 1 + image d'ombre 2) . Ceci nous permettrait d'élargir la taille de la base de données d'apprentissage (deux ou trois fois plus grande) et peut être vu comme une sorte d'augmentation de données (*data augmentation*) qui sera utile pour éviter le problème de surapprentissage (over-fitting) et améliorer la performance de notre méthode.

Notre thèse nous a permis de démontrer l'utilité de l'ombre projetée par une ou plusieurs sources lumineuses pour la reconnaissance de postures humaines élémentaires. L'ombre permettrait de fournir assez d'information complémentaire dans le cas de la présence d'une seule caméra afin de lever l'ambiguïté créée par la projection perspective.

BIBLIOGRAPHIE

Arduino. `http://arduino.cc/en/Main/arduinoBoardUno`, 2015.

Relayshield. `http://www.seeedstudio.com/wiki/Relay_Shield_V2.0`, 2015.

Shigeo Abe. *Support vector machines for pattern classification*, volume 53. Springer.

A. Abrams, K. Miskell et R. Pless. The episolar constraint : Monocular shape from shadow correspondence. Dans *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1407–1414, June 2013a.

A. Abrams, K. Miskell et R. Pless. The episolar constraint : Monocular shape from shadow correspondence. Dans *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1407–1414, June 2013b.

J. Ackermann, F. Langguth, S. Fuhrmann et M. Goesele. Photometric stereo for outdoor webcams. Dans *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 262–269, June 2012a.

J. Ackermann, F. Langguth, S. Fuhrmann et M. Goesele. Photometric stereo for outdoor webcams. Dans *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 262–269, June 2012b.

A. Agarwal et B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, Jan 2006a.

A. Agarwal et B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, Jan 2006b.

Ceyhun Burak Akgül. *Density-based shape descriptors and similarity learning for 3D object retrieval*. Thèse de doctorat, Ph. D. thesis, Dept. Signals-Images, Télécom ParisTech, Paris, 2007.

Edouard Auvinet, Franck Multon, Alain Saint-Arnaud, Jacqueline Rousseau et Jean Meunier. Fall detection with multiple cameras : An occlusion-resistant method based on 3-d silhouette vertical distribution. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):290–300, 2011.

A Bacelar, G GIMENEZ et al. 9-reconstruction 3d à partir des vues orthogonales d'un objet. 1994.

Bruce Guenther Baumgart. Geometric modeling for computer vision. Rapport technique, DTIC Document, 1974.

Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. Dans *Neural Networks : Tricks of the Trade*, pages 437–478. Springer, 2012.

James Bergstra et Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

Aaron F. Bobick et James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3): 257–267, 2001.

J. Y. Bouguet. Camera calibration toolbox for matlab. `https://www.vision.caltech.edu/bouguetj/calib_doc/`, 2008.

Jean-Yves Bouguet et Pietro Perona. 3d photography using shadows in dual-space geometry. *International Journal of Computer Vision*, 35(2):129–149, 1999.

Anusorn Bunteong et Nopporn Chotikakamthorn. Light source estimation using feature points from specular highlights and cast shadows. *International Journal of Physical Sciences*, 11(13):168–177, 2016.

Xiaochun Cao et M. Shah. Camera calibration and light source estimation from images with shadows. Dans *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 918–923 vol. 2, June 2005a.

Xiaochun Cao et M. Shah. Camera calibration and light source estimation from images with shadows. Dans *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 918–923 vol. 2, June 2005b.

German K. M. Cheung. Visual hull construction, alignment and refinement for human kinematic modeling, motion tracking and rendering. Rapport technique, 2003.

Chi-Wei Chu et I. Cohen. Posture and gesture recognition using 3d body shapes decomposition. Dans *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 69–69, 2005.

Dan Ciregan, Ueli Meier et Jurgen Schmidhuber. Multi-column deep neural networks for image classification. Dans *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

I. Cohen et H. Li. Inference of human postures by classification of 3d human body shape. Dans *2003 IEEE International SOI Conference. Proceedings (Cat. No.03CH37443)*, pages 74–81, Oct 2003.

Corinna Cortes et Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Thomas Cover et Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Tom Cuypers, Yannick Francken, Johannes Taelman et Philippe Bekaert. Shadow multiplexing for real-time silhouette extraction. Dans *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–68. IEEE, 2009.

D. E. DiFranco, Tat-Jen Cham et J. M. Rehg. Reconstruction of 3d figure motion from 2d correspondences. Dans *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–307–I–314 vol.1, 2001.

Lars Eldèn. *Matrix methods in data mining and pattern recognition*, volume 4. SIAM, 2007.

A. Fossati, M. Dimitrijevic, V. Lepetit et P. Fua. From canonical poses to 3d motion capture using a single camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1165–1181, July 2010.

Blender Foundation. Blender : free and open-source 3d computer graphics software. `http://www.blender.org/`, 1995.

Herbert Freeman. On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computers*, (2):260–268, 1961.

V. Girondel, L. Bonnaud, A. Caplier et M. Rombaut. Static human body postures recognition in video sequences using the belief theory. Dans *IEEE International Conference on Image Processing 2005*, volume 2, pages II–45–8, Sept 2005.

R. Gouiaa et J. Meunier. 3d reconstruction by fusioning shadow and silhouette information. Dans *Computer and Robot Vision (CRV), 2014 Canadian Conference on*, pages 378–384, May 2014a.

R. Gouiaa et J. Meunier. 3d reconstruction by fusioning shadow and silhouette information. Dans *2014 Canadian Conference on Computer and Robot Vision*, pages 378–384, May 2014b.

R. Gouiaa et J. Meunier. Human posture classification based on 3d body shape recovered using silhouette and infrared cast shadows. Dans *Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on*, pages 73–78, Nov 2015a.

R. Gouiaa et J. Meunier. Human posture recognition by combining silhouette and infrared cast shadows. Dans *Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on*, pages 49–54, Nov 2015b.

M. Grundmann, F. Meier et I. Essa. 3d shape context and distance transform for action recognition. Dans *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi et Kieran Greer. Knn model-based approach in classification. Dans *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer, 2003.

R. Guo, Q. Dai et D. Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, Dec 2013a.

R. Guo, Q. Dai et D. Hoiem. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2956–2967, Dec 2013b.

Lutz H Hamel. *Knowledge discovery with support vector machines*, volume 3. John Wiley & Sons, 2011.

Janne Heikkila et Olli Silvén. A four-step camera calibration procedure with implicit image correction. Dans *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 1106–1112. IEEE, 1997.

Henrique Steinherz Hippert, Carlos Eduardo Pedreira et Reinaldo Castro Souza. Neural networks for short-term load forecasting : A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55, 2001.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, February 1962.

Y. Iwashita, A. Stoica et R. Kurazume. People identification using shadow dynamics. Dans *2010 IEEE International Conference on Image Processing*, pages 2453–2456, Sept 2010a.

Y. Iwashita, A. Stoica et R. Kurazume. People identification using shadow dynamics. Dans *2010 IEEE International Conference on Image Processing*, pages 2453–2456, Sept 2010b.

Y. Iwashita, A. Stoica et R. Kurazume. People identification using shadow dynamics. Dans *2010 IEEE International Conference on Image Processing*, pages 2453–2456, Sept 2010c.

Yumi Iwashita, Adrian Stoica et Ryo Kurazume. Gait identification using shadow biometrics. *Pattern Recognition Letters*, 33(16):2148–2155, 2012a.

Yumi Iwashita, Koji Uchino et Ryo Kurazume. Gait-based person identification robust to changes in appearance. *Sensors*, 13(6):7884–7901, 2013.

Yumi Iwashita, Koji Uchino, Ryo Kurazume et Adrian Stoica. Gait identification from invisible shadows. Dans *SPIE Defense, Security, and Sensing*, pages 83711S–83711S, 2012b.

S. Ji, W. Xu, M. Yang et K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 221–231, Jan 2013. ISSN 0162-8828.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama et Trevor Darrell. Caffe : Convolutional architecture for fast feature embedding. Dans *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

Chia-Feng Juang et Chia-Ming Chang. Human body posture classification by a neural fuzzy network and home care system application. *IEEE Transactions on Systems, Man, and Cybernetics-part A : Systems and Humans*, 37(6):984–994, 2007.

I. N. Junejo et H. Foroosh. Using solar shadow trajectories for camera calibration. Dans *2008 15th IEEE International Conference on Image Processing*, pages 189–192, Oct 2008.

Adem Karahoca, Murat Nurullahoglu, M Demiralp, WB Mikhael, AA Caballero, N Abatzoglou, MN Tabrizi, R Leandre, MI Garcia-Planas et RS Choras. Human motion analysis and action recognition. Dans *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, numbéro 7. World Scientific and Engineering Academy and Society, 2008.

Ismail Khalid Kazmi, Lihua You et Jian Jun Zhang. A survey of 2d and 3d shape descriptors. Dans *Computer Graphics, Imaging and Visualization (CGIV), 2013 10th International Conference*, pages 1–10. IEEE, 2013.

Vili. Kellokumpu et Janne. Heikkilä. Human activity recognition using sequences of postures. Dans *Proc IAPR Conf*, pages 570–573. Machine Vision Applications, 2005.

Yeon Kim et J Aggarwal. Rectangular parallelepiped coding : A volumetric representation of three-dimensional objects. *IEEE Journal on Robotics and Automation*, 2(3): 127–134, 1986.

Stephan Kopf, Thomas Haenselmann et Wolfgang Effelsberg. Shape-based posture and gesture recognition in videos. Dans *Electronic Imaging 2005*, pages 114–124. International Society for Optics and Photonics, 2005.

Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Dans *Advances in neural information processing systems*, pages 1097–1105, 2012.

Jose-Luis Landabaso, Montse Pardàs et Josep Ramon Casas. Shape from inconsistent silhouette. *Computer Vision and Image Understanding*, 112(2):210–224, 2008.

Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994.

Yann LeCun, Léon Bottou, Yoshua Bengio et Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

EG McFarland, WW Carr, DS Sarma et JL Dorrity. Effects of moisture and fiber type on infrared absorption of fabrics. *Textile research journal*, 69(8):607–615, 1999.

Greg Mori et Jitendra Malik. Estimating human body configurations using shape context matching. Dans *European conference on computer vision*, pages 666–680. Springer, 2002.

Zouba. Nadia, Boulay. Bernard, Bremond. Francois et Thonnat. Monique. Monitoring activities of daily living (adls) of elderly based on 3D key human postures. Dans Barbara Caputo et Markus Vincze, éditeurs, *Cognitive Vision*, pages 37–50. Springer, Berlin, Heidelberg, 2008.

F. Nater, H. Grabner et L. Van Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. Dans *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2014–2021, June 2010a.

Fabian Nater, Helmut Grabner et Luc Van Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. Dans *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2014–2021. IEEE, 2010b.

K. Onishi, T. Takiguchi et Y. Ariki. 3d human posture estimation using the hog features from monocular image. Dans *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, Dec 2008.

Maxime Oquab, Leon Bottou, Ivan Laptev et Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. Dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Stefano Pellegrini et Luca Iocchi. Human posture tracking and classification through stereo vision and 3d model matching. *EURASIP Journal on Image and Video Processing*, 2008(1):1–12, 2007.

Bo Peng et Gang Qian. Online gesture spotting from visual hull data. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1175–1188, 2011.

Quoc-Cuong Pham, Laetitia Gond, Julien Begard, Nicolas Allezard et Patrick Sayd. Real-time posture analysis in a crowd using thermal imaging. Dans *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

M. Pierobon, M. Marcon, A. Sarti et S. Tubaro. Clustering of human actions using invariant body shape descriptor and dynamic time warping. Dans *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, pages 22–27, Sept 2005.

Mark W Powell, Sudeep Sarkar et Dmitry Goldgof. Calibration of light sources. Dans *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 263–269. IEEE, 2000.

Huimin Qian, Yaobin Mao, Wenbo Xiang et Zhiquan Wang. Recognition of human activities using {SVM} multi-class classifier. *Pattern Recognition Letters*, 31(2):100 – 111, 2010.

Parvin Razzaghi, Maziar Palhang et Niloofar Gheissari. A new invariant descriptor for action recognition based on spherical harmonics. *Pattern Analysis and Applications*, 16(4):507–518, 2013.

Jordi Sanchez-Riera, Jan Čech et Radu Horaud. Action recognition robust to background clutter by using stereo vision. Dans *European Conference on Computer Vision*, pages 332–341. Springer, 2012.

Silvio Savarese, Marco Andreetto, Holly Rushmeier, Fausto Bernardini et Pietro Perona. 3d reconstruction by shadow carving : Theory and practical evaluation. *International journal of computer vision*, 71(3):305–336, 2007.

Konrad Schindler et Luc Van Gool. Action snippets : How many frames does human action recognition require ? Dans *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

Steven A Shafer et Takeo Kanade. Using shadows in finding surface orientations. *Computer Vision, Graphics, and Image Processing*, 22(1):145–176, 1983.

M. Shinzaki, Y. Iwashita, R. Kurazume et K. Ogawara. Gait-based person identification method using shadow biometrics for robustness to changes in the walking direction. Dans *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 670–677, Jan 2015.

Partha Srinivasan, Ping Liang et Susan Hackwood. Computational geometric methods in volumetric intersection for 3d reconstruction. *Pattern Recognition*, 23(8):843–857, 1990.

Gaurav Srivastava, Johnny Park, Avinash C. Kak, Birgi Tamersoy et J. K. Aggarwal. *Multi-camera Human Action Recognition*, pages 501–511. Springer US, Boston, MA, 2014a.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever et Ruslan Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014b.

Xinghua Sun, Mingyu Chen et Alexander Hauptmann. Action recognition via local descriptors and holistic features. Dans *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 58–65. IEEE, 2009.

Makehuman Team. Makehuman : an open source tool for making 3d characters. `http://www.makehuman.org/`, 2000.

Matthew A Turk et Alex P Pentland. Face recognition using eigenfaces. Dans *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

H. Wang, A. Kläser, C. Schmid et C. L. Liu. Action recognition by dense trajectories. Dans *CVPR 2011*, pages 3169–3176, June 2011.

Juyang Weng, Paul Cohen, Marc Herniou et al. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on pattern analysis and machine intelligence*, 14(10):965–980, 1992.

N. Werghi et Y. Xiao. Recognition of human body posture from a cloud of 3d data points using wavelet transform coefficients. Dans *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 70–75, May 2002.

Wikipedia. Machine à vecteurs de support, 2016a. URL `https://fr.wikipedia.org/wiki/Réseau_neuronal_convolutif`. [Online ; accessed 31-12-2016].

Wikipedia. Machine à vecteurs de support, 2016b. URL `https://fr.wikipedia.org/wiki/Machine_à_vecteurs_de_support`. [Online ; accessed 27-12-2016].

Chen Wu et H. Aghajan. Model-based human posture estimation for gesture analysis in an opportunistic fusion smart camera network. Dans *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 453–458, Sept 2007a.

Chen Wu et Hamid Aghajan. Model-based human posture estimation for gesture analysis in an opportunistic fusion smart camera network. Dans *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 453–458. IEEE, 2007b.

Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

S. Yamazaki, S. Narasimhan, S. Baker et T. Kanade. Coplanar shadowgrams for acquiring visual hulls of intricate objects. Dans *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

Jason Yosinski, Jeff Clune, Yoshua Bengio et Hod Lipson. How transferable are features in deep neural networks ? Dans Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence et K. Q. Weinberger, éditeurs, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

Xi Zhang, Yanwei Fu, Shanshan Jiang, Leonid Sigal et Gady Agam. Learning from synthetic data using a stacked multichannel autoencoder. Dans *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 461–464. IEEE, 2015a.

Xi Zhang, Yanwei Fu, Andi Zang, Leonid Sigal et Gady Agam. Learning classifiers from synthetic data using a multichannel autoencoder. *arXiv preprint arXiv :1503.03163*, 2015b.

Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

Z. Zhou, E. E. Stone, M. Skubic, J. Keller et Z. He. Nighttime in-home action monitoring for eldercare. Dans *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5299–5302, Aug 2011a.

Zhongna Zhou, Erik Edward Stone, Marjorie Skubic, James Keller et Zhihai He. Nighttime in-home action monitoring for eldercare. Dans *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5299–5302. IEEE, 2011b.

J. Ziegler, Kai Nickel et R. Stiefelhagen. Tracking of the articulated upper body on multi-view stereo image sequences. Dans *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 774–781, June 2006.