Faculté des sciences et technologies -
Université de Lille

Licence 3 Informatique parcours Info / Miage / Math Info

Label Recherche a sein de l'équipe *FOX*

Reconnaissance d'émotions par réseaux de neurones à Spikes

Étudiant :
Alexandre Hulsken

Encadrants :
Pierre Tirilly
Benjamin Allaert

Janvier-Avril 2018

Sommaire

Introduction

Dans le cadre de ma dernière année de Licence Informatique à l'université des Sciences et Technologies de Lille, j'ai pu suivre cette option qui est le Label Recherche. Le présent rapport qui suit retracera donc le travail qu'il m'a été donné de fournir.

1.1. Environnement

Dans le monde, le nombre d'informations disponibles est immense. Et même si l'on ne se restreint seulement aux informations visuelles, il reste une abondante quantité de ces dites informations. Ce sont des données difficiles à analyser à cause de cette richesse (la palette de couleurs, les différentes textures, les mouvements possibles, les positions dans l'espace, la variabilité d'apparence de l'information, etc...), ce qui introduit ainsi diverses problématiques lorsque l'on souhaite les analyser.

C'est pourquoi l'équipe de recherche FOX (Fouille et indexation de dOcuments compleXes et multimédia) s'est positionné sur le problème des difficultés d'analyses de ces informations multiples. Ils travaillent ainsi sur de nouvelles techniques dans ce large domaine tel que l'étude de descripteurs de mouvement pour détecter et suivre les objets mobiles. Ils considèrent donc les quatre grands domaines de validations : le regard, l'événement l'émotion et la reconnaissance de personne réalisé à deux niveau d'échelles : l'individu et les flux.

Mon projet porte donc sur un domaine bien précis de ces différentes thématiques qui est la reconnaissance émotionnel d'individu.

1.2. Contexte

L'expression faciale est l'un des traits les plus important dans le domaine de la reconnaissance d'expression faciale[1]. Mais de nos jours automatiser ce travail et ne plus le laisser exclusivement aux êtres vivants est un réel défi. En effet, avec le déploiement et les avancées des agents conversationnels et de la robotique qu'il nous a été donné d'observer ces dernières années, leur apprendre à pouvoir faire cette même extraction d'informations pourrait permettre de pousser encore beaucoup plus loin les technologies déjà créées.

Ce travail de reconnaissance d'émotions consiste donc à inférer l'état émotionnel d'une personne (joie, peur, colère, etc...) à partir de signaux acquis par différents capteurs (visuels, audios, données physiologiques...). Dans ce projet nous nous intéressons à l'analyse des expressions faciales sur des images statiques.

Le passage du signal d'entrée à l'émotion est généralement réalisé à l'aide d'un algorithme d'apprentissage automatique. Dans ce projet nous nous intéressons à la détection des expressions faciales dans les images à l'aide de modèles d'apprentissage de type réseaux de neurones impulsionnels.

1.3. Problématique

Ces réseaux de neurones, plus proches du modèle biologique que les modèles classiques de type perceptron, sont prometteurs, mais leur fonctionnement est encore mal compris. Leur application à des problématiques réelles, comme la reconnaissance d'émotions, demande donc un travail important d'exploration des architectures de réseaux et de leurs paramètres, sur des sous-problèmes de difficulté croissante.

1.4. Objectifs

L'objectif est d'identifier un sous-problème et d'effectuer des expérimentations permettant d'évaluer les modèles de réseaux de neurones impulsionnels déjà utilisés dans l'équipe sur chacun de ce sous-problème. Le travail réalisé consiste donc à :

— identifier un sous-problème à traiter ;
— préparer les données correspondant à ce sous-problème ;
— effectuer les expérimentations ;
— évaluer et analyser les résultats obtenus.

1.5. Plan

Dans la suite de ce rapport vous pourrez trouver un état de l'art dans la section suivante. La Section 3, quand à elle, parlera du travail de prétraitement des images qui sera suivi par la phase expérimentale dans la section 4. Enfin, je conclurai dans la section 5.

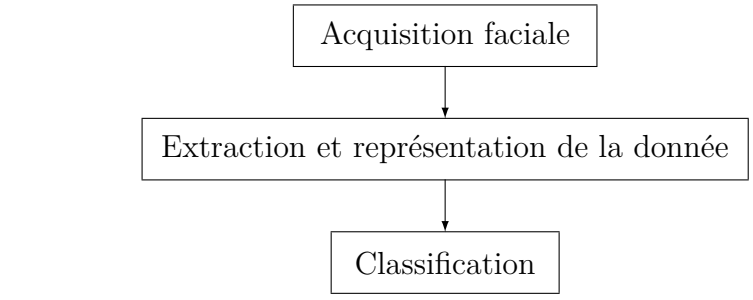
État de l'art

2.1. Processus d'analyse des expressions faciales

Ce sujet de recherche ayant eu un engouement particulier ces dernières années, il a été possible de voir l'émergence de différentes familles dans l'approche de l'analyse des expressions faciales.

2.1.1. Les méthodes traditionnelles

Le mécanisme de reconnaissance d'expressions faciales peut se décomposer en trois grandes différentes étapes[6] comme l'on peut le voire ci-dessous :



L'acquisition faciale, ou plus communément prétraitement de l'image, consiste à devoir trouver où se situe l'information que l'on va traiter dans l'ensemble de l'image et ensuite de la préparer pour la suite. La méthode la plus largement utilisé est d'extraire les points du visage. Pour les détecter, il existe différentes méthodes tel que les algorithmes de régression[4]. Ces différents points servent alors à savoir où se trouve certaines informations (tel que le nez, un coin de la bouche, les yeux, etc...) en fonctions de leurs coordonnées dans l'images et permettent donc par la suite d'avoir une estimation de la position de la tête du sujet dans l'espace et d'en extraire le visage que l'on souhaite traiter du reste de son environnement.

Dans les travaux qui ont été effectué sur l'extraction et la représentation des données, deux principaux modèles sont apparus. Le premier basé sur la géométrie fonctionne sur le fait d'extraire la forme ainsi que la localisation de partie du visage tel que la bouche, les yeux, le nez, etc... Ce modèle étant souvent couplé avec un FACS (Facial Action Coding System), qui permet de diviser une émotion en différents mouvements types ainsi que de leur intensité sur des zones spécifiques du visage tel que le plissement sur le haut du nez plus au moins fort, l'ouverture de la bouche ou un haussement de sourcil. Mais grâce aux

travaux de Shan, Gong et McOwan[10], l'importance de la qualité de l'image à pu être mise en évidence sur les résultats ce modèle alors qu'elle importe peu sur les méthodes d'extraction basé sur l'apparence (il faut noté que les travaux se basant sur la géométrie se base également sur l'apparence pour reconnaître les unités d'actions utiles pour le FACS). Cette dernière extrait l'information de l'émotion sur l'ensemble du visage après application de certains filtres. Mais dans chacun des cas, la représentation du la donnée extraite se fait par des vecteurs représentant chaque pixels de l'image que l'on souhaite traiter.

Une fois ce travail effectué, il ne reste plus que la classification. Selon Liu, Han, Meng et Tong[7], la phase de classification possède trois phases dans son apprentissage. La première étant l'acquisition de modèles suivi par une sélection de ces modèles qui a pour but de maximiser la différence entre les différences des descripteurs inter-classes et de minimiser les différences intra-classes. Suite à cela on passe enfin à la construction du ou des classifieurs qui permettront de classer leurs entrées en une seule et unique sortie correspondant à une émotion.

2.1.2. Les méthodes de deep learning

L'évolution dans le domaine de l'Intelligence Artificielle, et plus précisément dans les réseaux de neurones, a permis de faire d'énormes avancées dans le domaine de la vision artificielle. En effet, les réseaux de neurones convolutifs, de type MLP (MultiLayer Perceptron) grâce aux travaux de Lecun[5], utilisé dans la phase d'extraction de modèle a déjà permis de simplifier l'ensemble du processus de reconnaissance. Ils ont notamment simplifier le traitement des données puisque supportant des images brutes, il ne devient plus nécessaire de coder à la main l'ensemble de données sous certains formats, et il est donc possible d'entrer des images sous forme d'un vecteur de pixel. Ils ont également simplifier l'ensemble du traitement en fusionnant les étapes d'extraction et de classification (étapes 2 et 3) en une seule puisqu'ils sont capables de traiter les données et de retourner des entiers à leur sortie, ce qui correspond au label d'une classe (et dans notre cas à une émotion) tout ça en ayant de très bon résultats de réussite.

L'expérimentation de ces architectures d'apprentissage automatique se fait en séparant l'ensemble des données en deux sous-ensembles différents. L'un de ces derniers est donc utilisé pour l'apprentissage du réseau et le second pour son test. Ce cloisonnement a pour but de ne pas fausser les résultats de la phase expérimentale de l'architecture en ne faisant pas de test sur des images déjà apprises par le réseau.

2.2. Les Réseaux de neurones impulsionnels

Les réseaux de neurones impulsionnels ont une représentation similaire aux convolutifs. Ils sont constitués de neurones ayant plusieurs synapses d'entrés et un de sortie, et sont disposés également sous forme de couches. Mais malgré cela, ils possèdent un mode de très fonctionnement différents.

2.2.1. Le STDP neurone

Le neurone impulsionnel, à la différence de neurone convolutif, fonctionne grâce à une simulation d'impulsions électrique et la notion importante dans leurs utilisations est la notion de temps entre ces dites impulsions. Ils sont caractérisés par un ensemble de synapses (post et pré synaptiques) ainsi qu'un potentiel électrique, d'une fonction de déperdition et d'un seuil d'activation1. Leur fonctionnement étant que pour chaque impulsion pré-synaptique, la charge électrique du neurone augmente, et lors du temps où il n'est pas sollicité, cette charge diminue en fonction de leur fonction de déperdition. Ensuite lorsque cette charge dépassera le seuil du neurone, elle sera réinitialisé et le neurone enverra une impulsion post-synaptique.

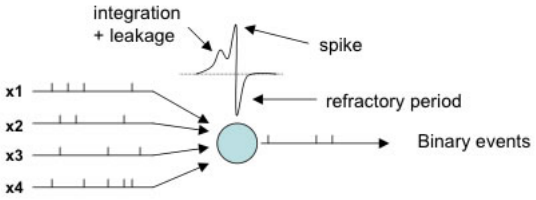


FIGURE 1 – Schéma d'un neurone impulsionnel

Il fonctionne également grâce à des poids sur ses synapses d'entrées, mais ici leur modification change de manière différente. Le principe sous-jacent est le même que celui du modèle biologique et se nomme la plasticité en fonction du temps d'occurrence des impulsions[9] (ou en anglais, Spike-timing-dependent plasticity, STDP). Il s'agit d'une règle d'apprentissage qui fonctionne en modifiant en regardant quand un synapse va emmètre une impulsion de sortie. Lorsque ce sera le cas, le principe sera d'augmenter le poids des synapses d'entrés ayant été activé avant l'émission de la sortie et va diminuer le poids de ceux qui seront activer après cette émission. Cette augmentation (ou diminution) du poids du synapse sera plus ou moins importante en fonction du temps entre l'impulsion d'entrée et celle de sortie (plus elles seront proches, plus la modification sera importante). Ce principe d'apprentissage du neurone est importante puisqu'elle implique une autre grande différence avec les modèles de réseaux de neurones traditionnels puisqu'il nous n'avons donc plus besoin de rétro-propagation pour l'apprentissage.

2.2.2. Une architecture particulière

Les réseaux de neurones impulsionnels possèdent une architecture assez particulière, puisque dans la plus grande majeur partie des cas, ne sont disposés qu'en une seule et unique couche. Une autre particularité est qu'en plus des synapses d'entrées/sorties, il existe également une deuxième catégorie de synapses appelé LIF[3] (Leaky-Integrate-and-Fire) qui relie chaque neurone avec le reste des autres. Ils ont pour but d'empêcher l'apprentissage du même modèle par plusieurs neurones. Ils sont chacun activé lorsqu'un neurone envoie une impulsion post-synaptique et lance une face inhibitrice des autres neurones durant laquelle ils laisseront leur potentiel électrique redescendre jusqu'à être nul.

Cette architecture de réseaux de neurones à Spikes sont plus proche du modèle biologique que les modèles classiques de deep learning. C'est pourquoi nous travaillerons dans la suite avec ce modèle d'architecture pour mon projet de recherche.

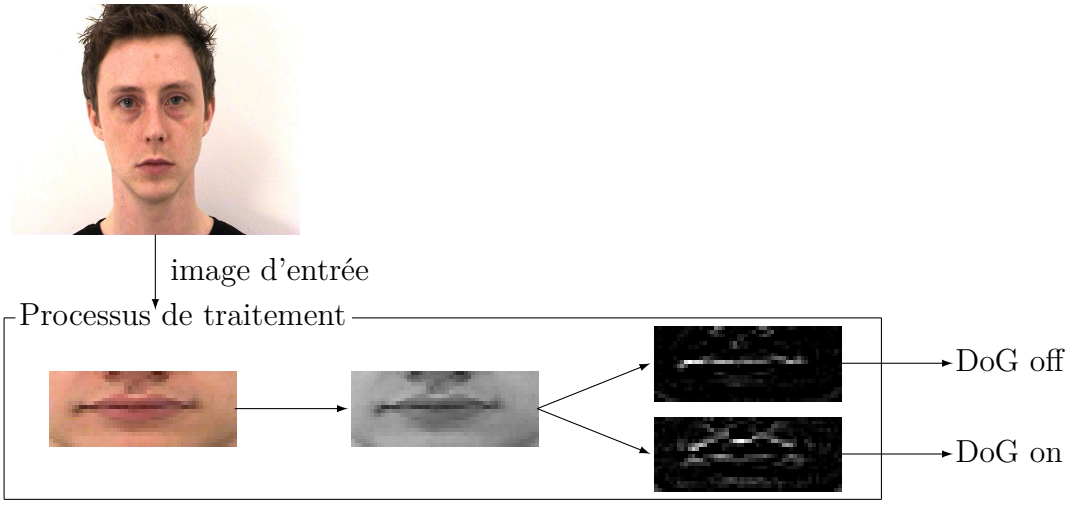


FIGURE 2 – Schéma du processus de prétraitement mise en place

Le prétraitement des données

La première partie de mon travail portait sur le prétraitement des images. La figure 2 représente le processus que j'ai implémenté avec une image extraite de la base de donnée ADFES.

3.1. La normalisation des données

Mon travail de prétraitement des données a ainsi commencé par extraire la partie importante du visage que je souhaitais utiliser. Pour cela j'ai donc utilisé les points du visage (un modèle à 68 points tel qu'ils sont représentés dans la figure 3) afin de pouvoir découper cette image grâce à la bibliothèque OpenCv et ne garder uniquement la partie qui me servirai pour la suite. J'ai également dû redimensionner l'image résultante en une taille prédéfinie afin de normaliser l'ensemble des images ce qui permettait par la suite de travailler avec le réseau de neurones à spikes.

3.2. L'extraction des contours

J'ai ensuite dû transformer ces images couleurs en noirs et blancs. Ce qui m'a ensuite permis de pouvoir appliquer un algorithme d'extraction des contours (dernière étape du

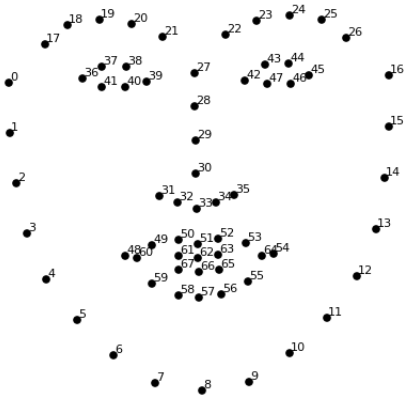


FIGURE 3 – Schéma représentant les points du visage

processus de prétraitement représenté dans la figure 2) fonctionnant sur un algorithme de différences de gaussiennes (DoG). Le principe fonctionne sur le fait que pour chaque pixel de coordonnées (x, y) on utilise un flou gaussien se modélisant sous la formule suivante :

$$G_\sigma(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2 + y^2}{2\sigma^2})$$

Où σ le paramètre du flou. Ensuite, il a fallu faire une soustraction pixel par pixel de deux flous paramétrées différemment (il est important que le paramètre σ soit différent pour ces deux flous, sinon l'extraction échouera). Finalement, la différenciation on/off du résultat se fait en ne retenant uniquement les valeurs positives(on) ou négative(off).

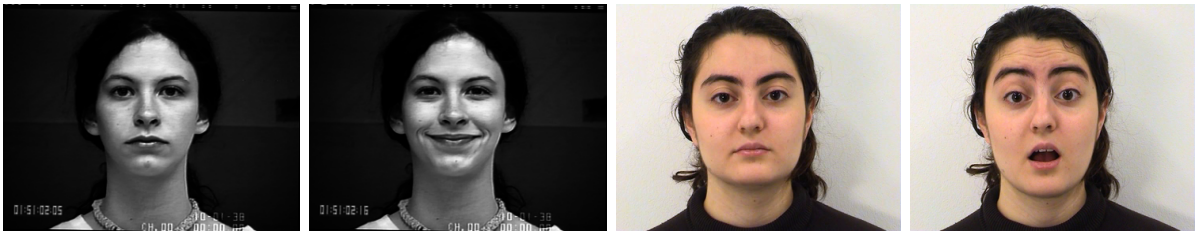


FIGURE 4 – A gauche se trouve deux images extraites de CK+ (représentant un visage neutre et souriant) et à droite deux autres extraites de ADFES (représentant l'émotion neutre et la surprise).

L'expérimentation

Il s'agissait maintenant, après avoir mis en place le processus de traitement de mes données, de tester l'architecture déjà utilisé dans l'équipe, pour cela nous allons voire quel à été ma démarche expérimentale.

4.1. Les bases de données

Pour l'expérimentation de l'architecture utilisé, nous avons utilisé la base de données CK+[8] (The Extended Cohn-Kanade database). Il s'agit d'une base très utilisé dans le domaine de la reconnaissance d'expressions faciales ce qui pourra nous aider dans la comparaison avec d'autres modèles déjà existant. Cette base étant constitué des 6 expressions basiques (la colère, le dégoût, la peur, la joie, la tristesse et la surprise) et d'une expression non basique (le mépris). Elle est ainsi divisée en 327 séquences d'images commençant par l'émotion neutre et allant jusqu'à l'apex, qui est le sommet de l'émotion, joué par 118 sujets.

J'ai également utilisé la base de données ADFES[2] (The Amsterdam Dynamic Facial Expression Set) ayant la même structure mais proposant un plus grand panel d'émotions non basiques (avec l'embarras et la fierté en plus). Tout cela joué par 22 sujets (10 femmes et 12 hommes).

Ces deux bases, représenté dans la figure 4, ont notamment des conditions d'acquisitions excellentes pour l'analyse puisque, comme étant faite en laboratoire, elles disposent des visages frontales dans une éclairage suffisant et homogène avec des expressions intenses et une bonne qualité de résolution. C'est pour cela que nous avons choisi d'effectuer les tests de l'architecture sur ces dites-bases.

Ensemble d'entraînement	ensemble de test	DoG	taux de reconnaissance
training set	testing set	on	100 %
training set	testing set	off	100 %
testing set	training set	on	99,4 %
testing set	training set	off	99,64 %
ADFES	CK+	on	74,2 %
ADFES	CK+	off	74,7 %
CK+	ADFFES	on	79,6 %
CK+	ADFFES	off	81,4 %

TABLE 1 – Tableau des premiers résultats expérimentaux

4.2. Le test de l'architecture du réseaux de neurones

Sur l'architecture que je souhaitais utiliser nous avons donc l'ensemble du prétraitement précédent pour l'acquisition (étape 1). Ensuite, nous avons donc Le réseaux de neurone pour l'extraction de modèle (étape 2) et nous utilisons une SVM (Support Vector Machine) pour la classification (étape 3).

4.2.1. Test de la difficulté du problème

Afin de tester la difficulté du problème que je mettais en place, et pour avoir une base de résultats afin de voire l'efficacité de l'utilisation d'un réseau de neurones impulsionnel, j'ai lancé une phase de test de l'architecture entrant les résultats du prétraitement directement dans la SVM (qui, comme le réseau de neurones, doit avoir une phase d'apprentissage) ce qui m'a permis d'obtenir les résultats que vous pourrez observer dans le tableau 1 (un schéma de l'architecture du test se trouve sur la figure 5 en annexe).

Dans cette phase de test j'ai donc mis en place deux protocoles de test différents, le premier étant un training/testing set où j'ai créé un ensemble en fusionnant les deux bases. J'ai ensuite extrait un nombre nombre prédéfini d'image de cette ensemble pour les mettre dans l'ensemble de test, le reste servait donc à l'ensemble d'entraînement. Le second protocole étant de la cross-dataset validation où l'on prend tout simplement une base pour l'apprentissage et l'autre base pour le test.

4.2.2. Le protocole expérimental

Au vu des résultats précédents, On a pu déduire que le problème était assez simple (puisque l'ensemble des résultats sont élevés). Mais ils ont également pu soulever des problèmes se

trouvant dans le premier protocole puisque les résultats sont beaucoup trop proches de la classification parfaite. Après analyse, il en est ressorti des problèmes dans la construction de ces ensembles de ce protocole, tel que le fait que les sujets pouvaient apparaître dans les deux ensembles, que plusieurs images était se ressemblaient au point d'être quasiment identiques et que certaine classe était en surpopulation. Il s'agit effectivement d'un problème falsifiant la fiabilité des tests puisque l'on testait l'architecture sur des images qu'il connaissait déjà.

Suite à ces premiers résultats j'ai fait le choix construire un protocole expérimental se basant sur la cross-dataset validation. J'ai donc fait un apprentissage du réseau de neurone sur une base. Suite à cela, il faut testé cette même base dans le réseau une fois qu'il était entraîné afin de pouvoir connaître le comportement qu'il a sur se qu'il connaît déjà afin de pouvoir utilisé se résultat comme base d'apprentissage pour la SVM. Finalement, il faudra testé l'architecture obtenu sur la deuxième base (vous pourrez trouver un schéma illustrant ce protocole se trouve en annexe sur la figure 6).

Conclusion

Dans le cadre de ce projet de recherche, j'ai travaillé sur la reconnaissance d'expression faciale d'un individu dans des images. Plus précisément, mes contributions se portaient sur deux parties différentes dans l'utilisation d'un réseau de neurones impulsionnels. D'une part, sur le processus de prétraitement des données en mettant en place un processus complet permettant d'extraire les contours d'une image sur la zone particulière de l'image qui nous intéresse. J'ai également pu travailler sur la mise en place d'un protocole d'expérimentation de ce dit réseau.

Nous pourrions donc envisager pour le future de ce projet d'augmenter la difficulté en plusieurs autres sous problèmes de la reconnaissance d'expression faciale en travaillant sur une autre zone du visage (comme les yeux ou le nez) ou sur plusieurs zone en même temps voir même sur le visage complet.

Remerciement

Je tiens tout particulièrement à remercier mes encadrants, Benjamin Allaert et Pierre Tirilly, pour m'avoir pris sous leurs ailes lors de ce Label Recherche. Je tenais à les remercier également d'avoir répondu à toutes mes questions et de m'avoir fait confiance lors de ce temps passé ensemble.

Je remercie aussi toute l'équipe FOX et l'IRCICA pour m'avoir accueilli chez eux.

Merci aussi à toute l'équipe pédagogique sans qui cette option ne serait pas là et aucun étudiant de Licence 3 n'aurait pu découvrir le monde de la recherche aussi tôt.

Je voudrais finalement dire merci aux amis suivants cette même option. Ils m'ont écouté et également posé des questions sur ce que je faisais avec intérêt ce qui m'a permis de trouver des solutions par moi-même à des problèmes et avancer encore plus loin.

ANNEXE A :

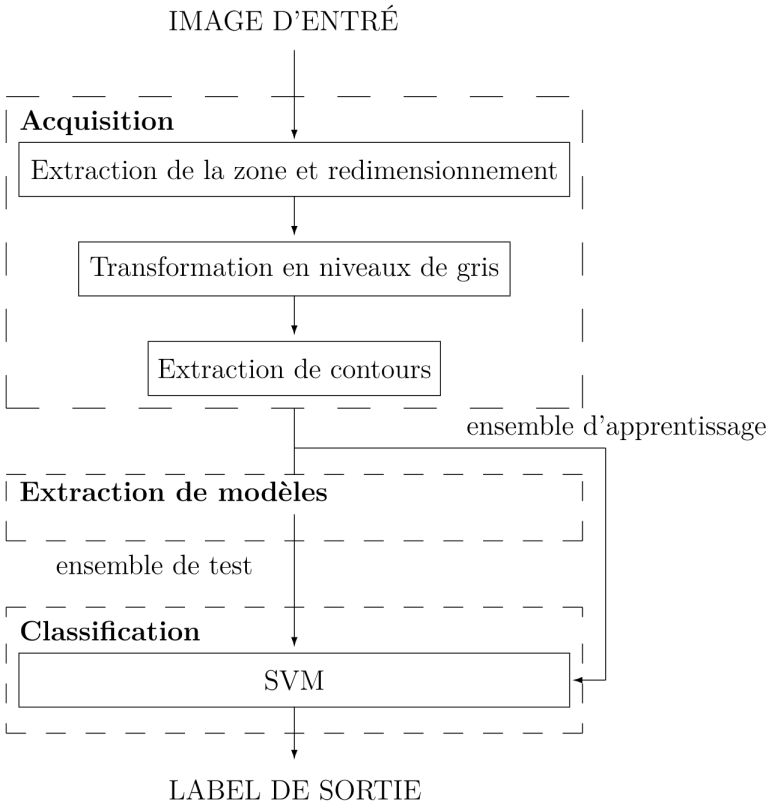


FIGURE 5 – Schéma de l'architecture pour le test de la difficulté du problème de test de l'architecture du réseau de neurones impulsionnels

ANNEXE B :

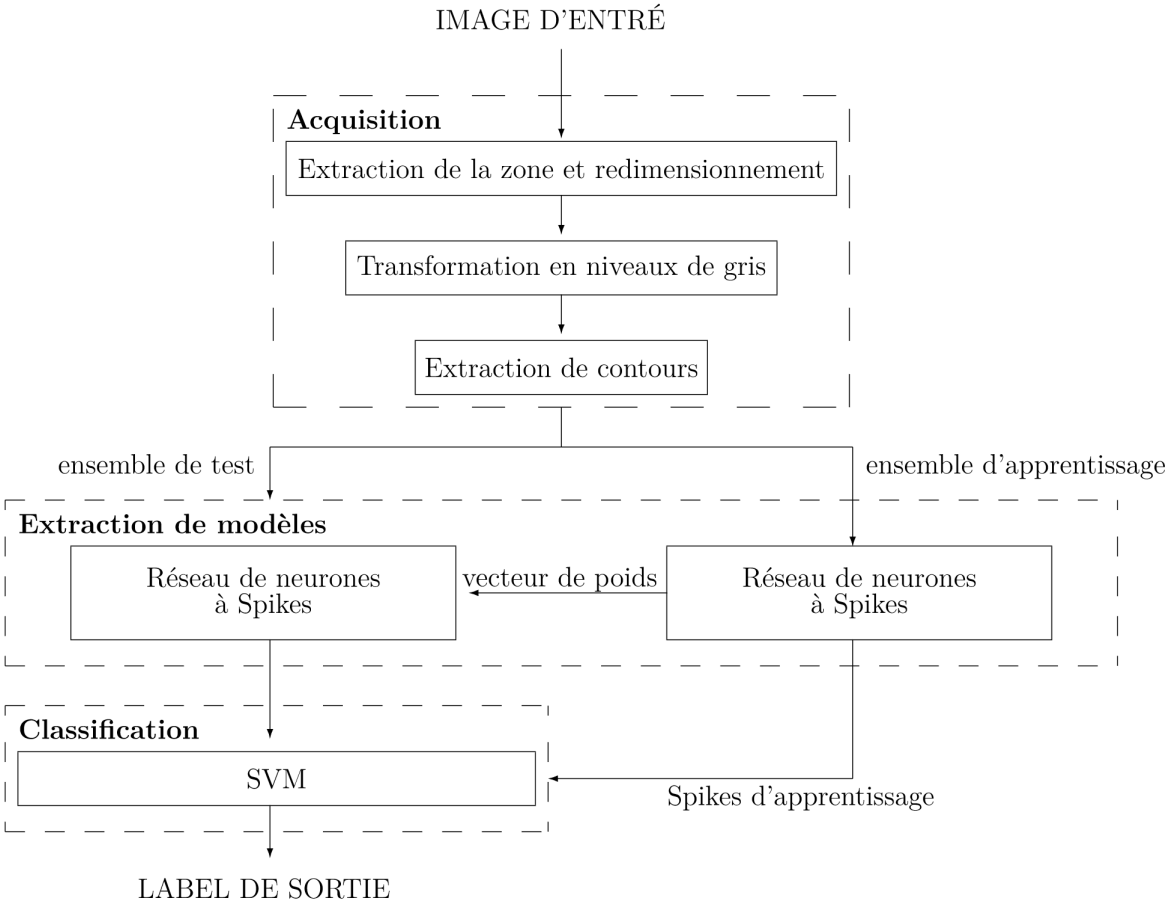


FIGURE 6 – Schéma de l'architecture utilisé dans ce projet

Références

[1] Charles Darwin. *The expression of the Emotions in Man and Animals*. CreateSpaec Independent, 2011.

[2] Job Van der Schalk, Skyler T. Hawk, Agneta H. Fischer, and Bertjan Doosje. Moving faces, looking places : Validation of the amsterdam dynamic facial expression set (adfes). 2015.

[3] Peter U. Diehl and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. 2015.

[4] Xijian Fan and Tardi Tjahjadi. A dynamic framework based on local zernike moment and motion history image for facial expression recognition. *Pattern Recognition*, 64 :399–406, 2017.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learn applied to document recognition. 1998.

[6] Stan Z. Li and Anil K. Jain. Handbook of face recognition. 2011.

[7] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. *Conference on Computer Vision and Pattern Recognition*, 2014.

[8] Patrick Lucey, Jerey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. 2010.

[9] Timothee Masquelier and Simon Jonathan Thorpe. Face feature learning with spike timing dependent plasticity. 2006.

[10] C. Shan, S. Gong, and P.W.McOwan. Facial expression recognition based on local binary patterns : a comprhensive study. *Image and Vision Computing*, 27, 2009.