Video Analysis for Micro- Expression Spotting and Recognition

Hua Lu

Thèse

UNIVERSITE BRETAGNE LOIRE

THESE INSA Rennes
sous le sceau de l'Université Bretagne Loire
pour obtenir le titre de
DOCTEUR DE L'INSA RENNES
Spécialité : Signal, Image, Vision

présentée par
Hua LU
ECOLE DOCTORALE : *MATHSTIC*
LABORATOIRE : *IETR UMR CNRS 6164*

Video Analysis for Micro-Expression Spotting and Recognition

Thèse soutenue le 05.04.2018
devant le jury composé de :

Christophe ROSENBERGER
Professeur, ENSI Caen / Président
Olivier ALATA
Professeur, Université de Saint-Etienne / Rapporteur
Mohamed DAOUDI
Professeur, Université de Lille / Rapporteur
Catherine SOLADIE
Maître de Conférences, CentraleSupélec / Examinateur
Pascal BOURDON
Maître de Conférences, Université de Poitiers / Examinateur
Kidiyo KPALMA
Professeur, INSA de Rennes / Directeur de thèse
Joseph RONSIN
Professeur, INSA de Rennes / Co-encadrant de thèse

Analyse de vidéo pour la détection et la reconnaissance de micro-expressions

Hua LU

En partenariat avec

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my two supervisors Prof. Kidiyo KAPLMA and Prof. Joseph RONSIN for the continuous support of my PHD study and related research, for their patience, motivation, and immense knowledge. Their guidances provided me many help in all the time of research and writing of this thesis. I could not have imagined having better advisors for my PHD study.

Besides, I would like to thank the rest of my thesis committee, Prof.Olivier ALATA, Mohamed DAOUDI, Christophe ROSENBERGER, Catherine SOLADIE and Pascal BOURDON, for their insightful comments and questions which encouraged me to widen my research from various perspective.

Furthermore, I am grateful to all those who gave me the possibility to complete this study. Especially, I would like to express my sincere gratitude to Prof. Mingqiang YANG, the supervisor of master, for his help, support and encouragement in overcoming numerous obstacles I have been facing. Meanwhile, I would like to thank Da CHEN, for the guidance in writhing skills and fruitful discussion with him. In addition, I am grateful for the scholarship provided by China Scholarship Council (CSC).

Nevertheless, I am also grateful to all my colleagues in Institut d'Electronique et de Télécommunications de Rennes (IETR) for providing me help and guidance both in my daily life and research work.

I would like thank my friends, HengYang WEI, Yi LIU, Lu ZhANG, Qiong WANG, Shishun TIAN, Fangyi CHAO, Lipin LI, Jingjun Gao and Ang Zhou, for their help and of course friendship.

Finally, I would like to thank my family : my parents, my sisters and brother for supporting me spiritually throughout writing this thesis and my life.

Résumé français

Chapitre 1 : Introduction

Les expressions faciales se produisent lorsque les gens ont tendance à exprimer leurs sentiments ou réagir à une situation. Dans notre vie quotidienne, les expressions faciales jouent un rôle important dans la communication avec les autres. En général, elles peuvent se diviser grossièrement en deux catégories : macro-expression et micro-expression. Contrairement aux macro-expressions qui durent longtemps et ont une forte manifestation sur le visage, les micro-expressions se caractérisent par un changement rapide qui dure moins d'une demi seconde et ont une faible intensité sur des parties du visage. Les micro-expressions apparaissent généralement dans les situations où les gens cherchent á contrôler ou gérer leurs émotions : une micro-expression, c'est ce qu'il reste lorsque la personne essaie de dissimuler une émotion. Généralement, dans la vie quotidienne, l'objectif principal du contrôle des émotions ou des mensonges est de lisser et faciliter les interactions sociales, de gagner l'estime et l'affection des autres. La politesse veut que l'on ne révèle pas les mensonges lorsqu'ils surviennent afin d'assurer un fonctionnement social normal. Cependant, il est important de détecter les mensonges commis avec une intention hostile. Les menteurs ne peuvent pas supprimer complètement toute expression faciale du visage et ainsi, la manifestation de micro-expressions peut servir comme un indice d'une tromperie. Ces propriétés inspirent les applications potentielles des micro-expressions dans les domaines de la sécurité, de la médecine, dans la détection de mauvaises intentions dans un aéroport, dans la perception de distorsions psychiatriques chez des patients, et dans l'efficacité des négociations dans les affaires.

L'objectif de l'analyse des micro-expressions dans des vidéos comprend leur détection et leur reconnaissance. Une grande variété d'approches de détection et de reconnaissance de micro-expressions a été exploitée au cours de la dernière décennie. Pour la détection de micro-expressions, une idée largement utilisée est de comparer les distances statistiques

entre des caractéristiques de la texture dérivées des différentes trames vidéo. Son principal avantage est sa faible complexité de calcul. L'utilisation de l'information de mouvement dérivée de la déformation du visage est un autre moyen de suivre les micro-expressions, par ex. la méthode de flux optique [SBF+14]. Cependant, les informations de mouvement restent souvent sensibles au bruit.

Pour la reconnaissance de micro-expressions, l'extraction des caractéristiques est l'un des points les plus importants. De nombreux efforts ont été déployés pour l'extraction de caractéristiques efficaces et discriminantes. Le motif local binaire (LBP-TOP), défini à partir de trois plans orthogonaux, est l'un des descripteurs choisis en premier pour la représentation des micro-expressions [PLZP11], en raison de sa capacité à décrire à la fois la forme et la dynamique des informations de texture des images du visage. Cependant, les méthodes basées sur LBP-TOP ont des difficultés à capturer les changements d'apparence subtils tels que les petites rides autour des yeux ou de la bouche, et par lesquels les résultats de la reconnaissance des micro-expressions sont dominants. Comme variantes, parmi les caractéristiques existantes, les caractéristiques de mouvement dérivées du flux optique sont choisies, en raison de leur capacité à caractériser, avec succès, des mouvements subtils sur le visage.

Dans cette thèse, deux méthodes de détection et une méthode de reconnaissance de micro-expressions sont proposées. Le reste de la thèse est organisé comme suit : le contexte de la thèse est introduit dans le chapitre 2, y compris les concepts de base sur les macro- et micro-expressions. Le chapitre 3 introduit plusieurs descripteurs de caractéristiques fondamentaux et le classificateur SVM (machine à vecteurs de support) utilisé dans la reconnaissance de micro-expressions. Deux méthodes proposées pour la détection de micro-expressions sont présentées dans le chapitre 4. Le chapitre 5 introduit les méthodes de reconnaissance de micro-expressions faciales basées sur des caractéristiques de mouvement. Enfin, la conclusion et des perspectives sont présentées dans le chapitre 6.

Chapitre 2 : Expressions faciales

Ce chapitre présente le contexte de cette thèse : les concepts de base sur les macro- et micro-expressions. La description de l'expression faciale et les différences entre macro- et micro-expressions ainsi que leurs relations sont présentées. L'outil classique d'analyse d'expressions faciales par système de codage d'actions faciales FACS (facial action coding system) faisant usage d'unités d'action (AU) analysant les expressions est présenté. Dans

ce chapitre, neuf bases de données sur les micro-expressions, construites ces dernières années, sont brièvement présentées.

Les macro-expressions sont des expressions faciales qui durent plus d'une demi-seconde et peuvent être facilement observées à l'œil nu tandis que les micro-expressions sont des expressions faciales qui se produisent dans un temps bref qui est beaucoup plus court que les macro-expressions. Une micro-expression révèle le vrai sentiment/émotion que les gens essaient de cacher ou d'inhiber. L'origine de la micro-expression peut remonter à l'hypothèse d'inhibition de Darwin [DP98] écrivant que certaines expressions faciales ne peuvent pas s'être créées spontanément mais en fait sont exprimées pour refléter l'émotion ressentie. Près de cent ans plus tard, Haggard et Isaacs [HI66a] ont rapporté trouver des micro-expressions lors d'un visionnage de films cinématographiques, image par image, dans une recherche sur l'échange thérapeute-patient. Quelques années plus tard, en s'appuyant sur les travaux antérieurs de Haggard et Isaacs, Ekman et Frisen [EF69] ont signalé l'existence de la micro-expression quand ils ont étudié les relations entre le mensonge et les comportements non verbaux du corps. La capacité à identifier les micro-expressions est une compétence importante pour lire l'émotion ressentie par une personne. Cependant, la plupart les gens ne peuvent remarquer l'apparition de micro-expressions, ni les reconnaître en temps réel. Elles apparaissent et disparaissent si vite que vous les manqueriez si vous cligniez des yeux. Aussi, les scientifiques tentent-ils de former les gens à reconnaître les micro-expressions. L'entraînement avec l'outil d'apprentissage de micro-expressions augmente le taux de reconnaissance des mensonges pendant lesquels des micro-expressions se produisent. Afin d'analyser les expressions faciales, le système de codage d'action faciale (FACS) basé sur les unités d'action faciales (AU) fut développé par Ekman et Friesen [EF76] pour décrire et distinguer les mouvements du visage. Une unité d'action est l'unité minimale du comportement facial, qui peut être combinée et prise en compte pour décrire toute expression faciale. Ils ont défini 44 AU qui sont présentées et étudiées en groupes en fonction de l'emplacement ou du type d'action en cause.

Pour l'étude des micro-expressions, neuf bases de données ont été établies dans les années récentes. Le tableau 1 résume brièvement ces bases de données.

Table 1: Résumé des bases de données de micro-expressions

Base		#Micro-Expressions	#Participants	Âge moyen	#Ethnies	Fps	Résolution	Elicitation	#Catégories d'émotion	FACS codé
Polikovsky		42	10	\	3	200	640×480	Posé	6	Oui
USD-HD		100	\	\	\	29,7	1280×720	Posé	6	Non
York-DDT		18	50	\	\	\	\	Spontané	\	Non
Canal9		24	195	\	\	\	720×576	Spontané	\	Non
CASME	A	195	35	22.3	1	60	1280×720	Spontané	7	Oui
	B						640×480			
SMIC	SMIC-HS	164	20	26,7	3	100	640×480	Spontané	3	Non
	SMIC-VIS	71	10			25				
	SMIC-NIR	71	10							
CASME II		255	35	22,03	1	200	640×480	Spontané	5	Oui
CAS(ME)^2		57	22	22,59	1	30	640×480	Spontané	4	Oui
SAMM		159	32	33,24	13	200	2040×1088	Spontané	7	Oui

Chapitre 3 : Extraction et classification des caractéristiques faciales

L'extraction et la classification de caractéristiques faciales sont les principales tâches dans un système d'analyse le micro-expressions. L'extraction des caractéristiques faciales et les méthodes d'apprentissage automatique sont toutes les deux essentielles pour obtenir une excellente performance de reconnaissance. L'extraction de caractéristiques est un processus de transformation de données brutes en vecteurs de caractéristiques qui décrivent correctement les données de sorte que la performance du modèle construit sur des données inconnues puisse être optimale. Une caractéristique descriptive est une représentation d'une image pour sa caractérisation. Ce processus implique l'extraction d'informations efficaces et l'ignorance des données non essentielles. Un bon vecteur de caractéristiques fournit des informations essentielles et discriminantes pour des tâches telles que la détection d'objets ou la reconnaissance d'images. Une fois ces vecteurs de caractéristiques obtenus, les vecteurs sont livrés à des classificateurs comme les SVM ou les forêts d'arbres décisionnels (ou Random Forest) pour produire la classification.

Une brève étude bibliographique est menée dans ce chapitre sur l'extraction des caractéristiques faciales et leur classification. Plusieurs méthodes fondamentales pour l'extraction de caractéristiques sont examinées et toutes constituent la base pour l'extraction de caractéristiques de micro-expressions. La plupart des travaux liés à l'analyse des micro-expressions impliquent l'amélioration ou la combinaison de ces méthodes fondamentales. Le principe de base du SVM est présenté, avec trois noyaux largement utilisés : à savoir le noyau linéaire, le noyau polynomial et le noyau gaussien RBF (radial basis function).

Chapitre 4 : Détection de micro-expressions

Les approches traitant de micro-expressions dans le domaine de la vision par ordinateur consistent à détecter et classifier les micro-expressions dans des vidéos. Cela inspire une série d'approches pour l'analyse de micro-expressions intégrant des techniques assistées par ordinateur. Ce chapitre présente deux axes basés sur deux nouvelles caractéristiques pour la détection de micro-expressions, qui seront présentées dans ce qui suit : la projection intégrale et la distance géométrique.

Détection de micro-expression à l'aide de la projection intégrale

Dans cette section, nous proposons une nouvelle méthode de détection des micro-mouvements en invoquant la projection intégrale (IP) [LKR17] comme caractéristique pour décrire les changements dans des blocs divisant l'image du visage. Fondamentalement, la nouvelle méthode consiste en une série d'opérations : le suivi du visage et son traitement, le recadrage et l'extraction des visages, le calcul de la distance χ^2 pour mesurer la dissimilarité entre la caractéristique de projection intégrale de chaque trame et celle des images de référence, et le seuillage et la détection de pics. Parmi ces procédures, la projection intégrale sera présentée en détail ci-après. De façon à réduire les effets du choix de l'image de référence, une nouvelle méthode de sélection d'une image de référence est développée. Elle convient à la tâche de détection dans de longues vidéos en sélectionnant automatiquement différentes images de référence. Par comparaison avec la méthode qui choisit toujours la première image comme référence, dans le cas d'une très longue vidéo, elle permet de réduire des erreurs s'accumulant le long d'une séquence.

Projection intégrale

En raison des difficultés pour les gens à lire les micro-expressions, il est nécessaire de trouver des méthodes appropriées pour capter les changements subtils et rapides du visage. L'approche IP est présentée dans ce qui suit et est considérée comme une technique utile pour l'extraction des traits du visage. Comme l'IP peut être extrêmement efficace pour déterminer la position des caractéristiques, Brunelli et al. [BP93] l'ont appliqué pour la reconnaissance du visage humain. Dans un travail récent [HZH+16a], une méthode combinant l'IP et le LBP a été choisie pour la reconnaissance de micro-expressions grâce à sa capacité à fournir la propriété de la forme des images faciales. L'IP est une méthode

simple et rapide d'extraction de caractéristiques qui peut réduire les caractéristiques d'une image 2D à une simple donnée 1D.

Soit $\Omega \subset \mathbb{R}^2$ le domaine de l'image et $I : \Omega \times D \to \mathbb{R}$ une séquence d'images de niveaux de gris, où $D \subset \mathbb{R}$ est l'espace temps. A chaque point $(x, y) \in \Omega$ et à l'instant t, la valeur d'intensité est notée $I(x, y, t)$ et la formule typique de la fonction IP peut être exprimée comme suit :

$$IP_t^H(x) = \frac{1}{y_2 - y_1} \int_{y_1}^{y_2} I(x, y, t)dy, \tag{1}$$

$$IP_t^V(y) = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} I(x, y, t)dx, \tag{2}$$

La prochaine section présente une autre approche de détection des micro-expressions utilisant les caractéristiques géométriques.

Détection de micro-expressions à l'aide de la distance géométrique

Dans cette section, une autre méthode de détection innovante est proposée en exploitant les distances géométriques (distances euclidiennes) entre des points clés définis sur un visage. Les distances euclidiennes entre les points clés peuvent capturer des déplacements subtils le long des séquences et se sont avérées convenir à différentes tâches d'analyse faciale. Comme le rognage et le recadrage des visages ne sont pas requis, une faible complexité de calcul peut être obtenue en comparaison avec d'autres méthodes d'extraction de caractéristiques de textures ou de mouvements. L'organigramme de la méthode proposée est résumé à la Fig. 1. L'algorithme initialement extrait 49 points clés du visage dans des séquences vidéo contenant une micro-expression dynamique, cela du début jusqu'à la fin en passant par l'apogée, en utilisant la méthode de descente supervisée [XT13]. Afin d'obtenir des localisations précises, il est important de recaler les points clés après que l'alignement est effectué. La tâche essentielle dans l'organigramme est "Distance géométrique (geometrical distance)", qui sera expliquée plus en détails dans ce qui suit. L'analyse de la distance est effectuée pour le calcul de la différence le long de la séquence vidéo et la caractéristique obtenue est fournie au SVM pour la classification de la séquence en micro-expression (ME) / en non micro-expression (Non ME).

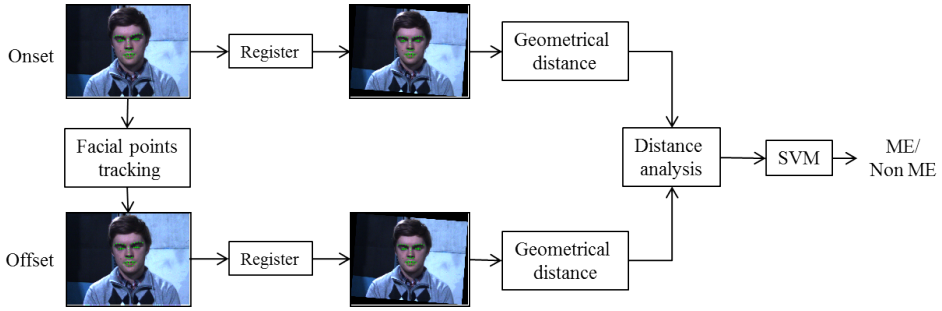


FIGURE 1: Schéma fonctionnel qui résume les différentes étapes de la géométrie faciale méthode.

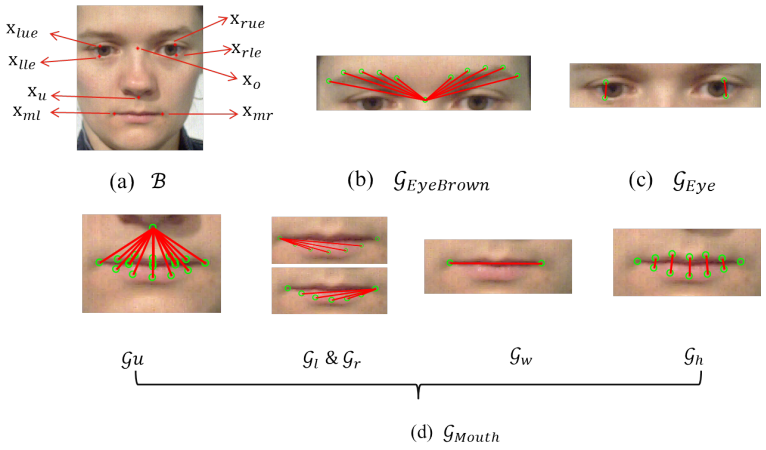


FIGURE 2: Une illustration des distances géométriques. (a) Emplacements des points de référence. (b)Distances géométriques entre les points des sourcils et le point de nez ($\mathcal{G}_b$). (c) Distances géométrique entre les points de paupières ($\mathcal{G}_e$). d) Les distances géométriques de la bouche ($\mathcal{G}_m$), y compris les distances entre les points de la bouche et le point sous le nez ($\mathcal{G}_u$), les coins des lèvres largeur($\mathcal{G}_l$ & $\mathcal{G}_r$), largeur de la bouche ($\mathcal{G}_w$) et hauteur de la lèvre ($\mathcal{G}_h$), respectivement.

Distance géométrique

Dans la méthode proposée, la classification est effectuée en fonction des informations de la géométrie des déplacements des points clés le long de la séquence vidéo, cela sans prendre en compte toutes les informations de texture du visage. Parmi les 49 points clés, les points du sourcil (10 points) et de la bouche (18 points) sont très sensibles à la plupart des expressions : voir Fig. 2. Il est rainsonnable de penser que la différence géométrique d'une micro-expression entre les images d'une séquence possède une dynamique le long de la séquence : elle commence à partir d'une petite valeur, puis atteint un pic et finalement retombe à une petite valeur.

Dans ce chapitre, deux méthodes de détection de micro-expressions sont proposées. L'-analyse sur les différences de la projection intégrale permet de détecter automatiquement les micro-mouvements avec une complexité de calcul faible. Les résultats expérimentaux sont positifs sur l'ensemble des bases de données CASME-A, CASME-B et CASME II, indiquant que cette méthode est capable de capter des micro-expressions dans des vidéos. À notre connaissance, c'est la méthode la plus rapide pour la détection automatique de micro-expressions et elle pourrait être mise en œuvre à l'avenir pour une détection en temps réel. La caractéristique géométrique est extraite de la face alignée, sans prendre en compte l'étape d'extraction du visage. La distance géométrique capture les petits changements pertinents plutôt qu'une caractéristique d'apparence. Ainsi, cette fonctionnalité est robuste à une variation d'éclairage. La performance sur quatre jeux de données de micro-expressions faciales (telles que SMIC-sub, SMIC-HS, SMIC-NIR, SMIC-VIS) démontre l'efficacité et le potentiel discriminant de la caractéristique géométrique. Au cours de l'expérience, il a été remarqué que les mouvements de la tête peuvent provoquer des détections erronées. Donc, á l'avenir, des algorithmes plus robustes devraient être étudiés pour résoudre ces problèmes.

Chapitre 5 : Reconnaissance de micro-expressions basée sur le mouvement

Ce chapitre présente les méthodes de reconnaissance des micro-expressions faciales basées sur des caractéristiques du mouvement. A la suite de l'examen des approches existantes pour la reconnaissance de micro-expressions, un nouvel opérateur, appelé fusion d'histogrammes des frontières de mouvement (FMBH), est proposé. Cet opérateur est

calculé en combinant d'une manière non linéaire les composantes horizontales et verticales du différentiel du flux optique. Il est établi en fusionnant les informations issues de ces champs de vecteurs externes. Les mouvements inattendus causés par un décalage résiduel apparaissant entre les images rognées de différentes trames sont supprimés et seul le mouvement relatif est capturé. Pour la construction des histogrammes respectifs, nous examinons également l'influence des différentes façons de construire une grille dense. Les études actuelles calculent les histogrammes des caractéristiques du mouvement en exploitant un nombre fixe de cases d'orientation spatiale. Notre travail étudie l'influence du nombre de cases d'orientation et de sa rotation de manière à tester la performance de reconnaissance. Afin d'extraire des caractéristiques discriminantes, la réduction de la dimensionnalité par la méthode de l'ACP est appliquée car elle présente des propriétés puissantes pour identifier la plupart des caractéristiques significatives et maintenir une forte corrélation entre deux caractéristiques de mouvement. Enfin, le classificateur SVM est utilisé pour la classification.

Fusion des histogrammes des frontières de mouvement : le descripteur proposé

Le mouvement facial peut être bien décrit en associant des caractéristiques de mouvement. Dans cette section, nous introduisons une nouvelle méthode de fusion de caractéristiques d'histogrammes de frontières de mouvement qui est établie sur la caractéristique de frontières de mouvement MBH. Les MBH sont calculés en termes, à la fois, des normes ou amplitudes $M_{\mathfrak{p}}$, $M_{\mathfrak{q}}$ et des angles $\theta_{\mathfrak{p}}$ and $\theta_{\mathfrak{q}}$. Nous définissons et considérons une fonction scalaire : $\alpha : \Omega \to S^1$ qui est calculé en termes de $\theta_{\mathfrak{p}}$ and $\theta_{\mathfrak{q}}$ par

$$\alpha(\mathbf{x}) = \mathfrak{E}_{\arctan}(\theta_{\mathfrak{q}}(\mathbf{x}), \theta_{\mathfrak{p}}(\mathbf{x})), \tag{3}$$

où $\theta_{\mathfrak{p}}$ et $\theta_{\mathfrak{q}}$ sont définis dans Eqs. (5.3) and (5.4). La fonction α peut être facilement utilisée pour établir un nouvel histogramme des orientations. Afin d'obtenir un histogramme pondéré, les normes $M_{\mathfrak{p}}$ et $M_{\mathfrak{q}}$ sont combinées ensemble. Cela peut être fait en considérant une fonction $M : \Omega \to \mathbb{R}$

$$M(\mathbf{x}) = \sqrt{\mathfrak{p}_x^2(\mathbf{x}) + \mathfrak{p}_y^2(\mathbf{x}) + \mathfrak{q}_x^2(\mathbf{x}) + \mathfrak{q}_y^2(\mathbf{x})} = \sqrt{M_{\mathfrak{p}}^2(\mathbf{x}) + M_{\mathfrak{q}}^2(\mathbf{x})} \tag{4}$$

pour tout $\mathbf{x} \in \Omega$. En fait, la fonction M dans Eq. (4) est la norme de Frobenius du jacobien de la matrice $\nabla\mathcal{F}$ du flux optique $\mathcal{F}$

$$\nabla\mathcal{F}(\mathbf{x}) = \begin{pmatrix} \nabla\mathfrak{p}(\mathbf{x}) \\ \nabla\mathfrak{q}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathfrak{p}_x(\mathbf{x}) & \mathfrak{p}_y(\mathbf{x}), \\ \mathfrak{q}_x(\mathbf{x}) & \mathfrak{q}_y(\mathbf{x}) \end{pmatrix}. \tag{5}$$

où $\mathfrak{p}_x(\mathbf{x})$ (ou $\mathfrak{q}_x(\mathbf{x})$) et $\mathfrak{p}_y(\mathbf{x})$ (ou $\mathfrak{q}_y(\mathbf{x})$) sont respectivement les d érivées de $\mathfrak{p}(\mathbf{x})$ et de $\mathfrak{q}(\mathbf{x})$ selon x et y. Nous avons maintenant obtenu la fonction d'orientation α et la fonction de pondération M. L'étape suivante consiste à construire le nouvel histogramme. Puisque les deux histogrammes utilisés dans MBH sont fusionnés ensemble, la méthode proposée peut être appelée "fusion d'histogrammes des frontières de mouvement" pour "fusion MBH" (FMBH).

Soit $\{\Theta_i\}_{1\leq i\leq P}$ une collection de sous-ensembles connectés de l'espace d'orientation S^1 satisfaisant $\Theta_i \cap \Theta_j = \emptyset, \forall i \neq j$ et $\cup_i \Theta_i = S^1$. Basé sur une telle partition, la fusion de l'histogramme des limites de mouvement $\mathcal{H}$ peut être construit avec un ensemble de fonctions caractéristiques χ_i

$$\chi_i(\mathbf{x}) = \begin{cases} 1, & \text{si } \alpha(\mathbf{x}) \in \Theta_i, \\ 0, & \text{autrement.} \end{cases} \tag{6}$$

et la fonction de pondération qui est la norme M telle que

$$\mathcal{H}(i) = \int_\Omega \chi_i(\mathbf{x})\, M(\mathbf{x})\, d\mathbf{x}. \tag{7}$$

La forme discrète $\hat{\mathcal{H}}$ de $\mathcal{H}$ peut être exprimée par

$$\hat{\mathcal{H}}(i) = \sum_{\mathbf{x}\in\mathbb{Z}^2} \chi_i(\mathbf{x})\, M(\mathbf{x}), \tag{8}$$

pour tout $\mathbf{x} \in \mathbb{Z}^2$, où $\mathbb{Z}^2$ est la grille de discrétisation orthogonale du domaine Ω.

La contribution principale, dans ce chapitre, réside dans la construction des caractéristiques faciales établies sur la base de champs de vecteurs de flux optiques différentiels. Dans ce but, un mappage non linéaire a permis d'établir une caractéristique fusionnant les champs de vecteurs de gradients respectifs des deux composantes du flux optique. Les caractéristiques proposées sont ainsi extraites des normes de Frobenius de la matrice jacobienne dérivée du flux optique. Pour évaluer et optimiser les performances des caractéristiques proposées, nous avons également étudié l'influence des différentes façons de construire les cases (bins) de l'histogramme en fonction du nombre de cases et de l'angle d'orientation initial de la première case. Les expériences menées sur quatre jeux de

données bien connus de micro-expressions montrent que la méthode donne des résultats prometteurs.

Chapitre 6 : Conclusion et perspectives

Cette thèse est consacrée à la détection et la reconnaissance de micro-expressions dans les vidéos. Nous avons proposé des méthodes d'extraction de caractéristiques pour l'analyse de micro-expressions. La contribution principale et les éventuels travaux futurs sont présentés dans ce qui suit.

Contributions :

- principe de détection des micro-expressions basé sur la fonctionnalité IP,
- principe de détection des micro-expressions basé sur la géométrie des points clé,
- principe de reconnaissance des micro-expressions basé sur une caractéristique de mouvement.

Travaux futurs :

- il serait utile d'étudier plus en profondeur la caractéristique géométrique pour la détection et la reconnaissance des micro-expressions,
- il vaudrait la peine de développer des approches plus puissantes d'alignement des visages,
- il serait utile d'explorer d'autres fonctions de mouvement, ou de combiner celles existantes,
- il serait intéressant d'étudier l'influence de différents classificateurs par apprentissage automatique pour la détection et la reconnaissance des micro-expressions,
- pour des travaux de recherche futurs dans ce domaine : plus de bases de données sur les micro-expressions devraient être construites à l'avenir. De nouvelles bases de données publiques et librement accessibles qui contiendraient plus d'échantillons, des données dans un véritable environnement de mensonge, des conditions variables d'occlusion, d'éclairage, etc. sont les bienvenues pour ce type de travail.

Contents

Chapter 1

Introduction

In our daily life, the facial expressions play a significant role in communication with others. People can convey their feelings by making facial expressions and can also know the emotions of others by reading facial expressions. In general, the facial expressions are roughly divided into two categories : the macro-expressions and the micro-expressions. In contrast to the macro-expressions which last a long time and have strong manifestation on face, micro-expressions can be characterized as a rapid change which last only less than a half of second and have a low intensity in parts of the face. Micro-expressions usually appear in situations where people want to control or manage their emotions. For most people, the main purpose of telling lies in their daily life is to smooth social interactions, or to gain the esteem and affection of other people. Politeness commands that we should not reveal these lies in order to ensure the normal social functioning. However, it can be important to detect lies with hostile intent. Liars cannot completely suppress facial expressions such that the manifestation of micro-expressions can be served as leakage or deception clue. These properties inspire the potential applications of micro-expressions in (1) high-stake situation, such as criminal investigations, airport and mass transit checkpoints, counter terrorism ; (2) business, including the sales, coaching, training, management, recruitment, leadership, business negotiations ; (3) medical treatment, such as doctor-patient consultation. However, despite the efforts of exploiting the micro-expression training tool by Ekman [Ekm02] for training people to recognize micro-expressions with naked eyes, only few experts have the ability of capturing micro-expressions by naked eyes. Given these difficulties suffered by the micro-expression detection by human, the requirements towards automatic facial micro-expression analysis

have insensitively increased in recent years. Even though the techniques in the field of computer vision/video understanding has been widely developed, the micro-expression analysis is still a challenging problem, since micro-expressions have the characters of brief duration and low intensity.

The goal of micro-expression analysis consists of two tasks : the detection and classification of micro-expressions in videos. A broad variety of micro-expressions detection and recognition approaches have been exploited in the past decade.

For micro-expressions detection, a widely used idea is to compare the statistical distances between texture features derived from different video frames, the main advantage of which is the low computation complexity. The approaches based on this distance-based idea take into account the basic features like the local binary patterns (LBP) [MZP14a] and the histogram of oriented gradient (HOG) [PKO09, DYL15] to characterize changes in the blocks of a face image. With these features in hand, the statistical distances such as the chi-squared distance can be used to compare differences between the reference frame and the successive frames. A micro-expression is made up of a set of successive frames, each of which has a distance value to the reference frame. Using the motion information derived from the face deformation is an alternative way to track micro-expressions, e.g. the optical strain method [SBF+14]. However, the motion information is sensitive to noise. A main directional maximal difference [WWQ+17] is proposed for solving this problem. Using the machine learning classifiers, more advanced detection methods have been devoted to this field, such as the temporal interpolation model [PLZP11], the re-encoded LBP using a re-parametrization of the second local order Gaussian Jet [RHP13], the random walk model [XFP+16] or the sparse sampling [LNSP17] combined with the support vector machine (SVM), multi-kernel learning or random forest classifiers. In their basic formulation, the supervised methods train a model to determine if a sequence does or does not contain a micro-expression. Although these methods achieve better detection performance than unsupervised models, the establishment of the training database will cost high computation burden.

For micro-expression recognition, the feature extraction is one of the most important steps. Many efforts have been done concerning the extraction of efficient and discriminant features. The local binary pattern from three orthogonal planes (LBP-TOP) operator is one of the firstly chosen descriptors for the representation of micro-expressions [PLZP11], due to its ability to describe both the shape and the dynamic texture information of face images. Furthermore, in order to improve the recognition results, a variety of the methods

that combine the LBP-TOP with other descriptors have been developed, including the LBP-STP [LHM+15], the combination of the Eulerian video magnification and the LBP-TOP [WRS+12], the STLBP-IIP model [HWL+16] and the STCLQP model [HZH+16a].

However, these methods based on the LBP-TOP have difficulties to capture the subtle appearance changes such as small wrinkles around the eyes or mouths, by which the recognition results of the micro-expressions are dominated. Alternatively, motion features derived from the optical flow are chosen due to their ability in successively characterizing subtle movements on face. The corresponding approaches include the main direction mean optical flow [LZY+16], the bi-weighted oriented optical flow [LSPW16] and the facial dynamics map [XZW17].

Besides the methods listed above, the approaches, such as the tensor independent color space model [WYL+14] and the combination of the local spatiotemporal directional features and robust principal component analysis in [WYZ+14], as well as the method based on the similar appearance of macro- and micro-expressions which aims to recognize micro-expressions by training a model in macro-expressions dataset [WSF11, JBY+17], also have obtained promising results.

Within this thesis, three methods for micro-expression detection and recognition are proposed. The main structure of the document is outlined as follows :

- **Chapter** 2 introduces the background : the basic knowledge for the macro- and micro-expression. We start this chapter from the description of the facial expression, and point out the differences between macro- and micro-expressions as well as their relations. Then the classical facial expressions analysis tool using facial action coding system (FACS) is introduced to make use of action units (AUs) for analyzing expressions. In this chapter, nine micro-expressions databases built in recent years are briefly introduced. Among them, three widely used databases including CASME [YWL+13a], CASME II [WJYWZ+14] and SMIC [LPH+13] are listed in detail. Specifically, the procedure of building databases is provided, including the way of expression elicitation, the way of selecting micro-expressions from raw videos and so on. The unsettled work in the establishment of micro-expressions databases is finally discussed.
- **Chapter** 3 introduces several fundamental feature descriptors and the support vector machine classifier used in micro-expressions recognition. Features include the LBP-based features, the optical flow-based features, the histogram of oriented gradient descriptor. Three kernels of the SVM are investigated.

- **Chapter** 4 introduces our two new methods for micro-expression detection.
  1. A micro-movement detection method exploiting the integral projection [LKR17] as a feature descriptor to characterize changes in the blocks of the face image is proposed. Basically, this new method consists of a series of operation : face tracking and processing, cropping and masking faces, integral projecting extraction, chi-squared distance computation for measuring the integral projection feature dissimilarity between each frame and the reference frame, and the thresholding and peak detection on obtained chi-squared distance values. In order to reduce effects of the reference frame choice, a new reference frame selection method is developed. It leads to the reduction of the errors accumulating along the sequence, when compared to method which always chooses the first frame. The proposed method is evaluated on two widely used datasets of CASME and CASME II through experimental comparisons with some popular feature extractors such as the OF, LBP and HOG operators. One of the main advantages of our method is its computation simplicity : the proposed method can obtain better or comparable results, but requiring much less computation time than the existing models using the OF, LBP and HOG operators.
  2. A novel detection method is proposed by exploiting the geometrical distances (Euclidean distance) on a face. The Euclidean distances between key points can capture subtle displacements along sequences and are proved to be suitable for different facial analysis tasks. Since the operation of cropping faces is not required, a lower computation complexity can be achieved in contrast with other texture or motion feature extraction methods. Experiments are conducted on the SMIC database by comparing the proposed method against state-of-the-art. The SMIC database consists of four sub-datasets which are SMIC-sub, SMIC-HS, SMIC-VIS, and SMIC-NIR. Comparative experimental results demonstrate that the proposed feature descriptor yields the best performance.
- **Chapter** 5 introduces facial micro-expressions recognition methods based on motion features. Upon the reviews of the existing approaches for micro-expression recognition, a new operator, called fusion motion boundary histograms (FMBH), is proposed. This operator is computed by combing both the horizontal and the vertical components of the differential of the optical flow. It is established by fusing

the information derived from these external vector fields in a nonlinear mapping manner. The unexpected motions caused by residual mis-registration that occurs between images cropped from different frames is removed such that the relative motion can be captured. For the construction of the respective histograms, we also examine the influence of different ways of building bins in a dense grid. Current studies compute the histograms of motion features by exploiting a fixed number of spatial orientation bins. This study investigates the influence of the number of orientation bins and its rotation so as to test the corresponding recognition performance. In order to extract discriminative features, the dimensionality reduction method of the PCA is applied since the PCA has powerful properties to identify most meaningful features and maintain a strong correlation between two motion features. Finally, the SVM classifier is employed for classification. The proposed feature is then validated and evaluated through the leave-one-subject-out (LOSO) protocol for micro-expression recognition. Moreover, the proposed method is compared to state-of-the-art methods on four well-known databases CASME, CASME II, SMIC and CAS(ME)^2. Comparative experimental results demonstrate that the proposed FMBH feature descriptor yields promising performance.

Chapter 6 summarizes the main contributions of this thesis and gives the perspective future work.

Chapter 2

Facial expressions

2.1 Introduction

Facial macro- and micro-expression analysis are topics of computer vision in recent years. Both macro- and micro-expressions play different roles in a human's life. They are facial expressions which mainly differ from each another regarding their lasting time. The studies associated to micro-expression begun half a century ago, and focused in the field of psychology. Finding solutions to micro-expression related problems by computer vision has attracted more and more attention thanks to the establishment of a broad variety of available micro-expression datasets.

This chapter begins by giving general concepts involving facial expressions that contain facial macro and micro-expressions in Section 2.2. The facial action coding system (FACS) in Section 2.4 is specially presented, which breaks down facial movements into a number of action units (AUs) and describes the facial changes in terms of AUs. This coding system can be exploited as a useful tool for building macro- and micro-expression database. Next, several micro-expression datasets are described in Section 2.5. The conclusion is presented in Section 2.6.

2.2 Facial Macro-Expressions

Macro-expressions are facial expressions which last more than half second and can be easily observed by naked eyes. Most studies associated to facial expression analysis refer to the macro-expression analysis. Facial expression is one of the most powerful channels of communication for human beings to convey their feelings, intentions, personality and

so on. As a non-verbal behavior, the facial expression plays an important role in our daily life. People can express their feelings by making facial expressions and can also communicate with others by reading facial expressions. During the past two decades, facial expression analysis has been paid huge attention in many fields. Psychologists studied the human psychology conveyed by the changing facial expression and computer scientists relied on the digital process to analyze the expression.

As one of the most famous pioneers for scientific exploration of facial expression, Darwin [DP98] hypothesized, based on his theory of evolution, that certain facial expressions of emotion appeared to be universal across various countries and culture. For a long time, the universality of facial expressions has remained standing debates in the psychology. Some emotion theorists carried out researches in various cultures to demonstrate that there exists several universal (basic or fundamental) emotions using the same elemental of facial muscle movements across the world [Tom84, EF86, EFO$^+$87, Bro91, MF92, CH92]. For example, Tomkins et al. [Tom84] proposed 8 basic emotions, fear, anger, joy, sadness, disgust, acceptance, surprise and curiosity. Ekman et al. [EF86] reported 7 basic facial expressions of anger, fear, surprise, sadness, disgust, contempt and happiness. However, there exists arguments against this theory. The early opinion is culture specific view, that what a facial expression implies is different from culture to culture [LaB47]. Ortony [OT90] questioned the assumption in theory of basic emotions and the view that which emotions are basic ones. Russell [Rus94] also doubted the reliability of the basic emotions concept which is plausible proposed by western psychologists and concluded that the association between facial expressions and emotion labels may vary in widely different cultures.

Despite the controversy, the universality theory of facial expression is widely accepted by many researchers and has had a profound impact on the modern automatic facial expression analysis. However, according to Russell [Rus94], contempt and disgust have been found to be confused with each other such that contempt has aroused the most controversy and is thought to be the least well established of the basic emotions proposed by Ekman and Friesen [EF86]. Thus, researchers have been focusing on developing automatic recognition systems that recognize the six basic expressions except contempt. Automatic facial expression recognition has drawn an increasing attention within the computer vision in recent years due to its wide applications in many areas such as human-robot interaction [ZG00] and computer facial animation [PW08]. Fig. 2.1 illustrates samples of seven basic facial expressions from Cohn-Kanade database (CK) [KCT00] and its ex-

tended database (CK+) [LCK+10]. Each expression is labeled by the action units (AUs), which will be introduced in Section 2.4.

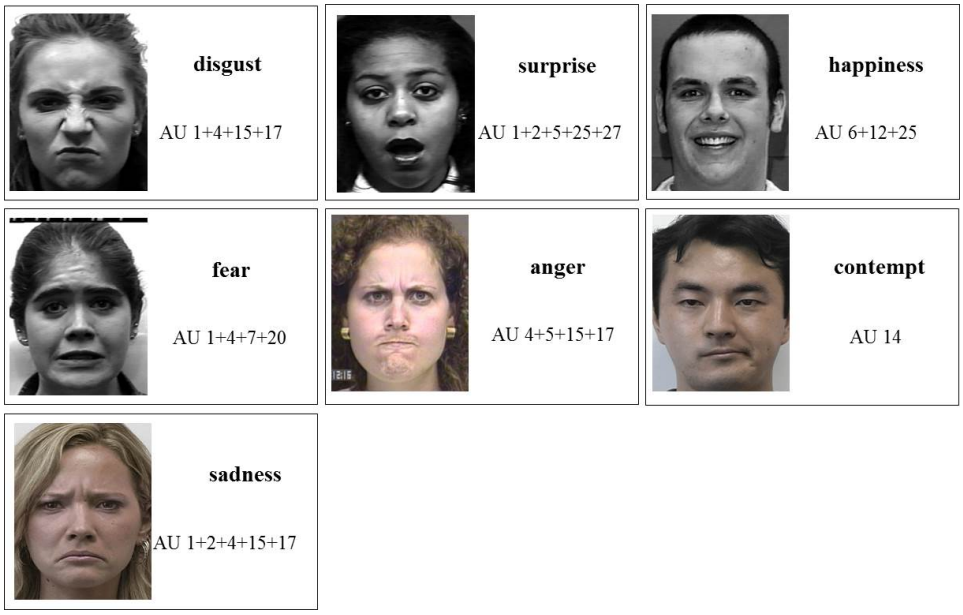


FIGURE 2.1: Examples of seven basic facial expressions from Cohn-Kanade (CK) database. The gray images are from the original CK database and the color images are samples from the extended database (CK+). Each expression is labeled by the action units (AUs), which will be introduced in Section 2.4.

2.3 Facial Micro-Expressions

The concept of "micro-expression" is proposed by Ekman and Frisen [EF69] and widely accepted by researchers. Micro-expressions are facial expressions that occur within a brief time which is shorter than macro-expressions. A micro-expression is a facial expression, which reveals the true feeling that people try to hide and suppress.

Micro-expressions occur when people are trying to manage their facial expressions purposefully. There are three major ways in falsifying a facial expression [EF75] : (1) an expression is simulated so that you express an emotion with an expression while you feel nothing, (2) neutralized so that nothing is manifested on your face while indeed you do have a particular feeling, or (3) masked so that a felt emotion is covered by other emotions corresponding to different expressions. A micro-expression is the result of interruption, which occurs in deintensifying, neutralizing, or masking a felt facial expression [EF75] (Page 163). When you sense from your facial muscles that you are

beginning to show the true feeling, you try to conceal the expression appearing on your face by deintensifying, neutralizing or masking it. The felt emotion remains on the face within a fraction of a second, named micro-expressions, followed immediately by a falsified facial expression.

There exists two features distinguishing micro-expressions from macro-expressions.

Short duration. The duration of the micro-expression is the main feature which differentiates from macro-expression. Obviously, the duration of the micro-expression is variable. However, the definition of the micro-expression's duration still remains ambiguous. The range of the micro-expression is firstly defined as '$40 - 200$' ms in the study of Ekman and Friesen [EF75], followed by '333 ms or less' in [ER97] and 'less than 250 ms' in [Ekm09]. Recent researchers provide different versions of definitions. Matsomoto and Hwang [MH11] consider the micro-expression ranges from 67 to 500 ms. Shen et al. [SWF12] suggest that a proper upper limit of duration of micro-expressions may be around 200 ms based on their experiments' observations. Yan et al. [YWL$^+$13b] conduct the study to provide evidence to define micro-expression by duration and present a duration of $170 - 500$ ms. Based on these observations, the duration of the micro-expression should be further discussed.

Low intensity. Micro-expressions involve a subtle movement. They may be very brief full (entire face), partial (only occurring in specific area) or slight (not much muscular contraction) expressions [Ekm07]. Figure 2.2 illustrates seven micro-expressions with AUs labels, where the AUs are depicted in Section 2.4. Comparing with macro-expressions in Figure 2.1, micro-expressions involve low intensity of the facial muscles.

Regardless of whether they are macro or micro, facial expressions are dynamic temporal process that match the time and duration of facial deformations and are described with three important features : onset, apex and offset [Ekm09, Bet12]. The onset is the point at which the expression starts to show up, the apex is the instant when the deformation of the expression reaches a peak and the offset represents the instant when the expression fades away. Hence, the micro-expression detection is a temporal segmentation of videos, which includes locating the micro-expression appearance instant and providing the duration between the onset and offset. Fig. 2.3 present an image sequence of the micro-expression which is labeled by 'disgust' in CASME [YWL$^+$13a]. A subtle "nose wrinkling" with the label of AU9 can be observed in the third picture. Due to the limit space, only five frames are presented including 1st(onset), 5th, 9th (apex), 13th and 16th (offset).

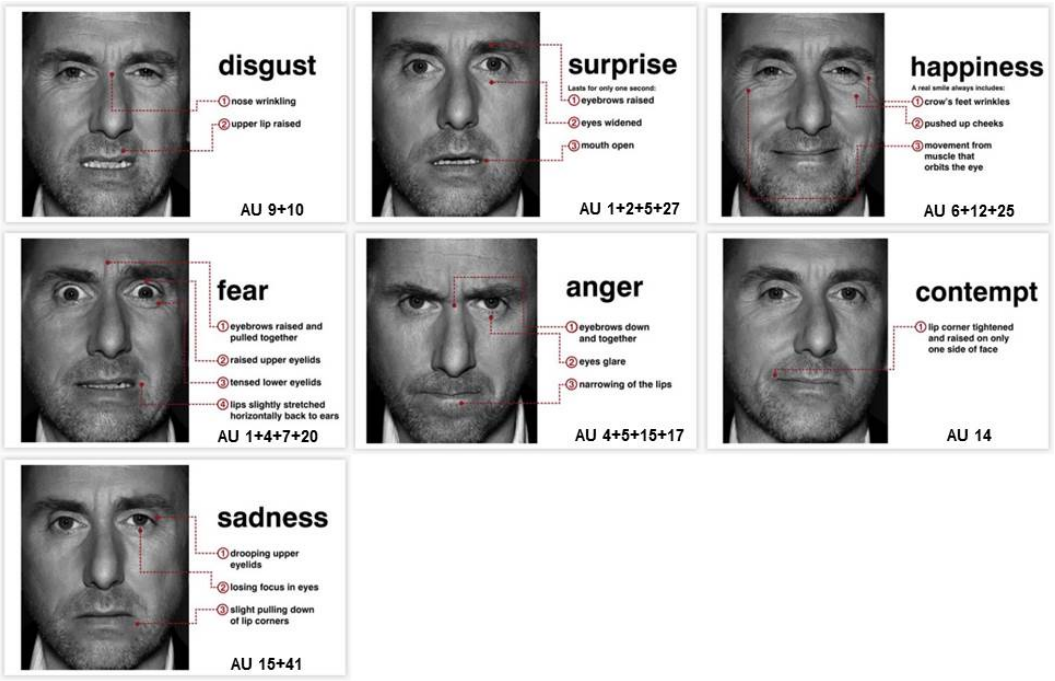


FIGURE 2.2: An example of seven micro-expressions. Each micro-expression is labeled by the action units (AUs) with detailed interpretation. Figure reprinted from [Dan10]

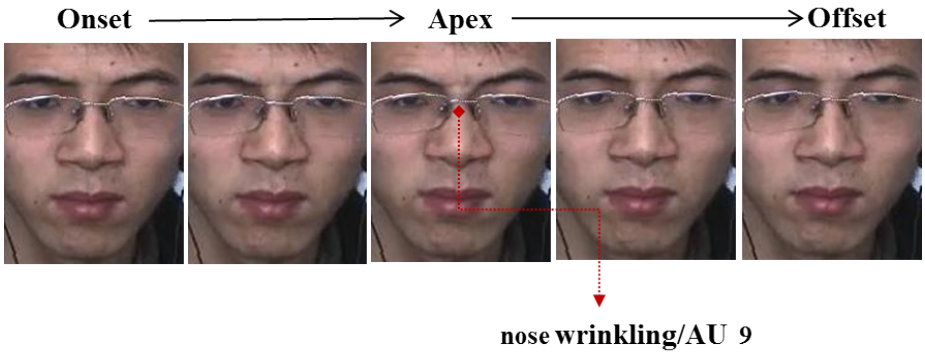


FIGURE 2.3: An example of a micro-expression sequence. Five frames are presented including 1st(onset), 5th, 9th (apex), 13th and 16th (offset).

The origin of the micro-expression can be traced back to Darwin's inhibition hypothesis writing that certain facial expressions can not be spontaneously created but in fact are expressed to reflect the felt emotion. In his book [DP98], he wrote :

A man when moderately angry, or even when enraged, may command the movements of his body, but...those muscles of the face which are least obedient to the will, will sometimes alone betray a slight and passing emotion. (Page 79).

Nearly one hundred years later, Haggard and Isaacs [HI66a] reported to find micro-expressions when scanning motion films frame-by-frame in therapist-patient interchange research. They watched occasionally that the expression on patient's face changed dramatically within a few fraction of second. They named this expression as "micromomentary expression (MME)" and found that the MME is meaningful to therapeutic process. In the study of Haggard and Isaacs, the MME could be spotted with a slow motion projection and could not be seen at normal rate.

A few years later, building on the earlier work of Haggard and Isaacs, Ekman and Frisen [EF69] reported the existence of the micro-expression when they studied the relations between deception and nonverbal behaviors of the body. They indicated that : (1) as one of nonverbal behaviors, the micro-expression may serve as leakage or deception clue ; (2) micro-expressions occur on part of face, in other words, they may be fragments of a squelched, neutralized, or masked display ; (3) the appearance of micro-expressions may be similar with macro-expressions but may be greatly reduced in time ; (4) micro-expressions can be read by expert clinical observers without the benefit of slow motion projection, but most people without training have difficulty to detect micro-expressions ; (5) proper training could help people improve the ability to recognize micro-expressions.

The study of micro-expression has inspired some researches concerning the measurement of individual difference in emotion recognition ability (ERA) which varies across gender, ethnicity, culture, and psychiatric status [MLWC$^+$00]. In order to measure this ability, Ekman and Friesen [EF74] developed the Brief Affect Recognition Task (BART) which involves a brief facial presentation (under 200 ms). Another measurement study was conducted by Matsumoto and his colleagues [MLWC$^+$00], who explored a new test named the Japanese and Caucasian Brief Affect Recognition Test (JACBART). In the JACBART, seven basic facial expressions (anger, contempt, disgust, fear, happiness, sadness and surprise) were utilized and displayed briefly (under 200 ms). Each facial expression was embedded in the middle of a 1000 ms presentation of the same poser's neutral expression.

Although the micro-expressions have been noted since Darwin, Porter and Ten Brink [PTB08] were the first to verify the existence of micro-expression with a scientific study, in which the genuine and falsified facial expressions of emotion were investigated. Their study showed that : (1) partially supporting the Darwin's inhibition hypothesis theory to prove that the occurrence of inconsistent expressions were observed more frequently in masked than in genuine expressions. However, genuine neutral expressions and neutralized expressions of felt emotion cannot be distinguished by inconsistent expressions ; (2) questioning the assumption that micro-expressions may occur on entire face, proposed by Ekman [EF75]. In fact, experiments in this study pointed out that micro-expressions were partially appeared, only in the upper or the lower face ; (3) partial supporting micro-expressions as a cue to deception. Their experiments pointed out that the micro-expressions occurred in the both genuine and falsified (including simulated, masked and neutralized [EF75]) emotional contexts. These partial micro-expressions were reliable indicator of the deception when they occurred in masked expressions. However, their occurrence in genuine expressions provides implications of false-positive errors, leading to the questionable applications such as airline-security (potential human-rights violations).

The ability to identify micro-expressions is an important skill for reading a person's felt emotion. However, most people cannot recognize the micro-expressions, not even notice their occurrence in real time. They appear and disappear so fast that you would miss them if you blink. Thus, scientists attempt to train people for recognizing micro-expressions. One is led by Ekman [Ekm02] and his group [1], named micro-expression training tool (METT), which aims at training people to recognize by naked eyes seven basic micro-expressions (anger, contempt, disgust, fear, happiness, sadness and surprise). The alternative tool is the micro-expression training videos (METV) [2], developed by Wezowski and his colleagues [WM16] for detecting micro-expressions on faces filmed on videos. Training with these tools increases the accurate recognition rate of deceptions in which micro-expressions occur.

The existence of micro-expression is relevant to deception in daily life. In the studies of DePaulo et al. [DKK+96], lying was defined as "intentionally try to mislead someone". Most people tell lies in every day life to smooth social interactions, to gain the esteem and affection of other people [DK98]. Politeness commands that we should not try to reveal these lies when they occur to ensure the normal social functioning. But it is

1. https ://www.eiagroup.com/us/training/online-training/
2. http ://www.microexpressionstest.com/

important to detect lies told by individuals with hostile intent. The study led by Hurley and Frank [HF11] says liars cannot completely suppress facial expressions such that the manifestation of micro-expressions can be regarded as the deception clues.

The existence of micro-expression is related to lies in high-stakes contexts such as criminal investigations, airport and mass transit checkpoints, counter terrorism, and so on [HF11]. For example, if someone intends to pose a threat to airline passengers, and is transiting a security checkpoint, he may have a fear of discovery. In order to hide the true feeling, he will try to manage his emotion resulting in more subtle manifestation of facial expressions. These micro-expressions can be detected by experts. In fact, the Transportation Security Administration (TSA) of United States launched a program called Screen Passengers by Observation Technique (SPOT), which is designed to identify people who could pose a threat to airline passengers, depending on 94 signs of stress, fear, or deception. There are about $3,000$ of TSA officers working at some 161 airports across the United States. The foundation of this programme is based on the micro-expression's study developed by Ekman and his group [Wei10]. Detecting these lies has assumed public safety and national security.

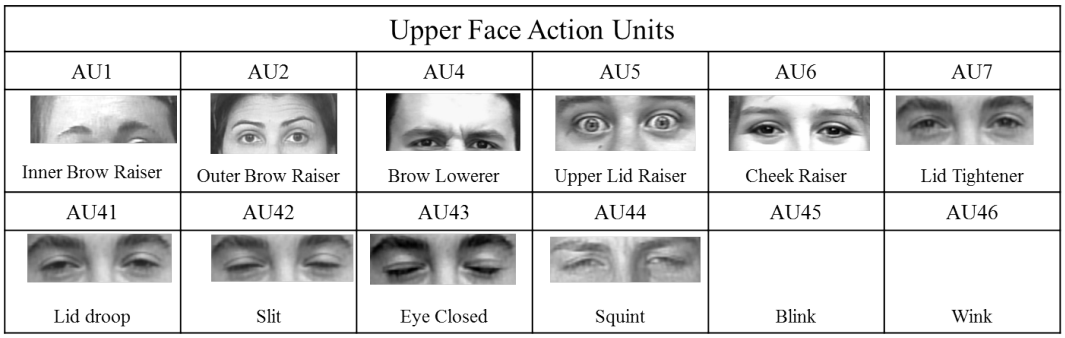
According to [WW12], the recognition of micro-expression is useful in business, including the sales, coaching, training, management, recruitment, leadership, business negotiations. Research in companies led by Wezowski and his wife Kasia [WW12] prove that the best sales people and negotiators are experts in reading body language and micro-expressions. It is a huge advantage in business if people can see what somebody feels.

Other applications of micro-expressions involve medical treatment, such as doctor-patient consultation. For clinicians, being capable of perceiving facial expression may aid in interpreting how much pain a patient is experiencing. However, some clues towards individuals who would repeatedly attempt suicide are most likely to be missed by doctors if they do not directly state the emotional impact on the patient. These clues are the manifestation of micro-expressions. The clinicians' person perception accuracy is related to the ability to recognize micro-expressions on patients. The study in [EL09a] proved that training with the METT improve the recognition of the static facial micro-expression (use one image instead of a sequence) of those medical students identified as good at communication with patients.

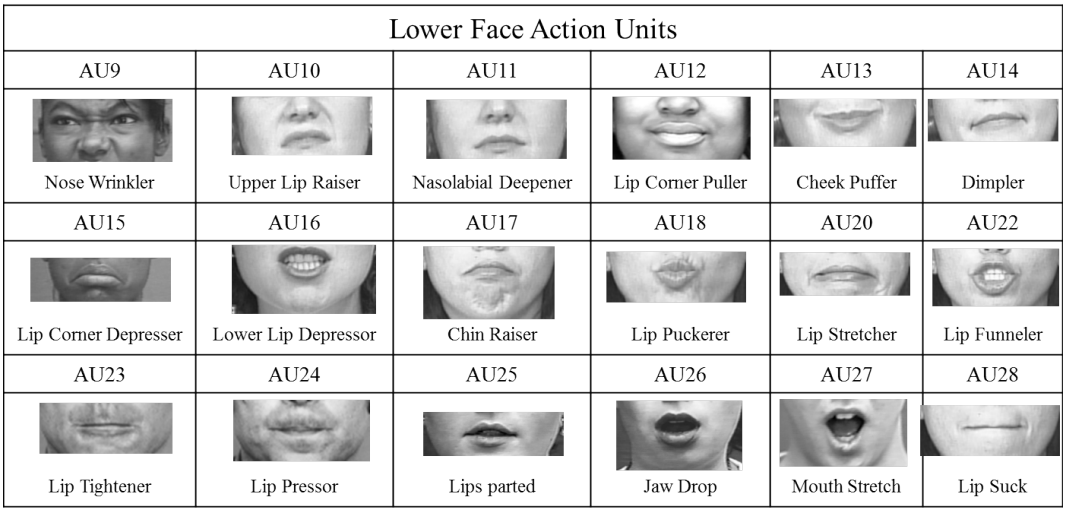
In the following, the facial action coding system (FACS) will be introduced, which is one of the most important tools for facial expression analysis.

2.4 Facial Action Coding System (FACS)

The previous researches on facial behavior studies depend on observers to infer from the whole face leading to inaccurate inferences about emotions and different interpretations. In order to solve this problem, Ekman and Friesen [EF76] developed a facial action coding system (FACS) to describe and distinguish facial movement based on action units (AUs). An action unit is the minimal unit of facial behavior, which can be in combination accounting for any facial expression. They defined 44 AUs which are presented and learned in groups based upon the location or type of action involved. AUs are divided into two main groups based on their location : AUs of the upper face and AUs of the lower face. The upper and lower face AUs are illustrated in Fig. 2.4.



(a)



(b)

FIGURE 2.4: **(a)**The upper face AUs. **(b)** The lower face AUs. Figure from [CAE07].

AUs involve facial muscles. Part of the action units are only associated with one facial muscle, and some action units are controlled by several muscles. For instance, AU 15 is the action of lip corner depressor which is controlled by Depressor anguli oris muscle. AU 26 represents the action of chin raiser which involves Mentalis muscle, see Fig. 2.5 for an intuitive display of facial muscles.

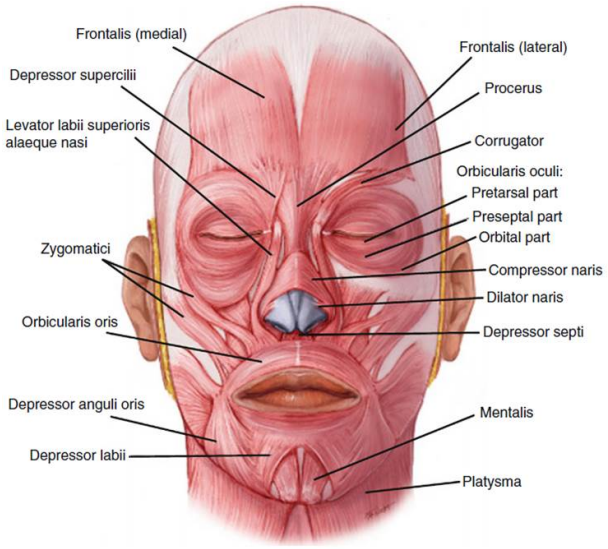


FIGURE 2.5: Facial muscles. Reproduced from [Pre13].

AUs can occur singlely or in combination. In [Sch85], more than 7000 different combinations of AUs have been observed. The combinations of AUs can produce relatively independent changes in appearance, changes in which one action masks another, or a new and distinctive set of appearances. For example, the appearance changes for AU $1 + 2$ are a sum of the appearance changes caused by AU 1 and AU 2 independently, without AU distorting or masking the appearances of the other. Another combination of the AU $1 + 2 + 4$ produces an appearance which is not simple the sum of the appearance of the individual AUs, but creates new and distinctive appearance. Examples of the seven expressions and their AU labels are shown in Figure 2.1.

The FACS provides an effective facial expression analysis tool so that any complex expression can be analyzed by breaking it down into a series of motion units. Thus, the FACS is considered to be a useful tool in micro-expression analysis, especially in the establishment of the micro-expression databases, which will be introduced in the following.

2.5 Micro-Expression Databases

The well-established database is the foundation for developing micro-expression detection or recognition system. Building a database that satisfies the different requirements and will be widely used for testing new algorithm is a difficult and challenging task. For micro-expression recognition, it poses various challenges in terms of building a standardized database, including the way of expression elicitation, the way of selecting micro-expression from raw videos. One of the most important problem is that expressions can be posed or spontaneous. Posed expressions are the artificial expressions that a person yields when someone asked him or her to do so. It usually happens when the subject is under observation in laboratory. In contrast, spontaneous expressions are the ones that people produce spontaneously, when people are involved with natural conversations, watching films etc. Posed micro-expressions are easy to capture and recognize, while spontaneous expressions are difficult to be produced and selected. They are different in appearance and temporal dynamic. Developing micro-expression analysis system implies that the spontaneous rather than posed expressions are recognized. Thus, current researchers have started focusing on building spontaneous micro-expression databases and developing spontaneous expression analysis. According to [Bet12], "*a standardized training and testing database contains images and video sequences (at different resolutions) of people displaying spontaneous expressions under different conditions (lighting conditions, occlusions, head rotations, etc)*". Nine micro-expressions databases were built in recent years. Three of them are widely used in micro-expression analysis : (i) the Chinese Academy Of Sciences Micro-expression (CASME) [YWL$^+$13a] ; (ii) the Spontaneous Micro-expression Database (SMIC) [LPH$^+$13] and (iii) the improved CASME (CASME II) [WJYWZ$^+$14]. Following paragraphs will present these nine databases in detail.

Polikvsky's Database [PKO09] contains 10 university student subjects (5 Asian, 4 Caucasian, 1 Indian) who were asked to perform 7 basic emotions with low facial muscles intensity and to go back to the neutral face expression as fast as possible, trying to simulate the micro-expression emotion. They used high speed camera with 480×640 resolution at a frame rate of 200fps. The drawback is that all the micro-expressions in this database are mimic expressions instead of spontaneous ones.

USF-HD [SGGS11] contains 100 micro-expressions and 181 macro-expressions. Videos of 47 sequences lasting on average approximately 1 minute in length were collected by cameras at a resolution of 720×1280 and a frame rate of 29.7 fps. Instead of sponta-

neous facial expressions, subjects were asked to mimic example videos containing micro-expressions to present micro-expressions. Out-of-plane head motion was avoided in order to decrease the difficulties of detecting micro-expressions.

York Deception Detection Test (York-DDT) [WSB09] includes fifty participants (31 females and 19 males) who are students of University of York except one administrator, all are native English speakers, aged between 18 and 45. It was first recorded as part of a psychological study for a deception detection test (DDT) at 25fps. All the participants were instructed to deceive or tell the truth when describing an emotional clip of a surgery or non-emotional film clip of a sunny beach. When they observed a surgical procedure, they were asked to describe it as if watching a beach scene, whereas if they saw the beach scene, they were asked to describe it as if watching a surgical procedure. In total, the deception detection task (DDT) consisted of 20 videos which had an overall length of 23 min and 15 seconds and each clip varied between 46 and 85 second with a mean length of approximately 60 seconds. 18 micro-expressions [PLZP11] were found on faces of 9 participants (3 male and 6 female) : 7 from the emotional and 11 from the non-emotional scene. There existed only 7 frames in the shortest clip.

Canal9 [VDFS09, SGGS11] comprises 70 political debates recorded by the Canal9 local TV station, which is used for analysis of social interactions and micro-expression research. Canal 9 database is recorded in HD format with a resolution of 720×576 and for a total of 43 hours and 10 minutes, where 19.7% of data time are recorded in the camera view of the full group (all people involved), 66.1% of data time is in personal shots, and 11.0% of data time is in multiple participants. Fig. 2.6 shows those three frequent camera views. Each debate was around a main question with a yes/no answer like " Are you favorable to the new laws on scientific research ? " In terms of gender, twenty−five women and one hundred and sixty-five men are involved in these debates. These videos were annotated according to socially relevant features (turn-taking, agreement−disagreement, and role) and low level descriptors (speaker segmentation, shot segmentation and so on). There were twenty−four sequences containing a micro-expression (each around 6 seconds) [SGGS11]. The advantage of this database is that all the micro-expressions are spontaneous and reflect real emotions that politicians try to hide. It is a challenging task to spot micro-expressions in this dataset due to other irrelevant facial movements and non−frontal camera view.

The Chinese Academy Of Sciences Micro-expression (CASME) [YWL$^+$13a] consists of two classes : the CASME-A and CASME-B, which were recorded by BenQ

FIGURE 2.6: Most frequent camera views in Canal9 dataset. Figure is reprinted from [VDFS09]

M31 camera with the resolution set to 1280×720 and Point Grey camera with resolution of 640×480, both filmed at 60fps. Moreover, the CASME-A and B were built in different lighting conditions, where participants from CASME-A were recorded in natural light while the samples in class B were filmed in a room with two LED lights that brought about uneven illumination on face. Participants (13 females, 22 males, the mean age of 22.03 years) were asked to seat in front of the 19−inch monitor with a camera on a tripod and watch the video episodes with high emotional valence for recording their emotions. Neutral faces are retained before and after the occurrence of each micro-expression. For coding process, two well-trained coders comprehensively went through the recordings and selected all micro-expressions that were no more than 500 ms or onset duration less than 250 ms because fast-onset facial expressions were considered as micro-expressions.

Some participants showed no micro-expressions. 195 short videos of 19 participants of 35 were selected, lasting 0.2 -11.7 s with an average time of 3.2 s. Most of videos contain one micro-expression sequence with the labeled onset, apex and offset. These micro-expressions are labeled by AUs based on FACS coding system [EF77]. The CASME-A includes 95 micro-expressions of 7 subjects, and the CASME-B consists of 100 expressions of 12 subjects. For all subjects, the number of micro-expressions range from 2 to 38. The criteria for classifying an emotion depends on the video content, self-report of participants. Eight classes of micro-expressions were captured including disgust 46, surprise 21, happiness 9, fear 2, contempt 2, sadness 6, repression 40 and tense 71. Unlike seven basic facial expressions, the tense and repression emotions that frequently appear in daily life are defined and introduced as micro-expressions in this dataset. The dataset provides micro-expressions sequences as well as videos for detection and recognition. Fig. 2.7 illustrates some samples from the CASME. In fact, an image sequence consists of more than five frames. Three key frames (onset, apex and offset) are labeled in this figure. Each

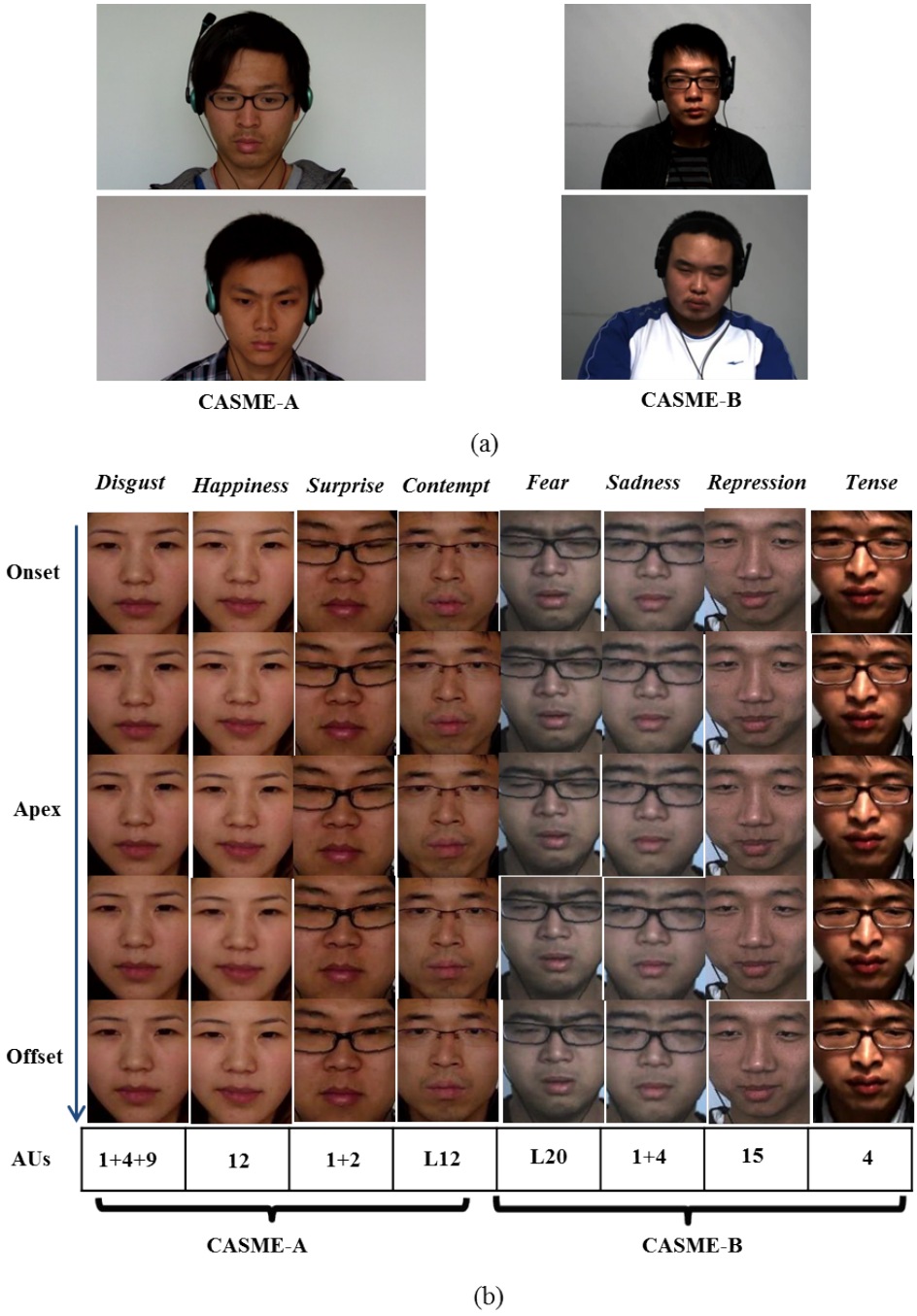


FIGURE 2.7: Samples from the CASME dataset. (a) Raw images from videos. Images of first and second column are samples from the CASME-A and CASME-B, respectively. (b) Cropped images. From left to right : *disgust*, *happiness*, *surprise*, *contempt*, *fear*, *sadness*, *repression* and *tense*. We provide five frames of each expression from the onset to offset labeled with AUs. $1-4$ columns are cropped samples of the CASME-A and the rest belongs to the CASME-B.

sequence is labeled as action units.

Spontaneous Micro-expression Database (SMIC) [LPH+13] was established by the Oulu institution and has three versions according to the released time. The first version of SMIC (referred as SMIC-sub) [3] only includes 77 video sequences of 6 subjects data [PLZP11] recorded by a high speed camera of 100 fps. Lately, they released the second version of SMIC which contains three subsets, including the SMIC-HS, SMIC-VIS, and SMIC-NIR, which were recorded separately by a high speed (HS) camera of 100 fps, a normal visual camera (VIS) and a near-infrared (NIR) camera at 25 fps, all of three with the resolution of 640×480. However, the first version and second version only consist of short video clips rather than long video clips. In recent times, an extended version of SMIC (refered as SMIC-E) was released and was also divided into three datasets of the SMIC-E-HS, the SMIC-E-VIS and the SMIC-E-NIR, in which the SMIC-E-HS includes 13 subjects data of 157 long video clips, the SMIC-E-VIS and SMIC-E-VIR both contain 8 subjects data of 71 long video clips of average duration of 5.9 seconds.

For the second version, 20 participants with a mean age of 26.7 attended the recording experiment, among those 6 were females and 14 males, 10 Asians, 9 Caucasians and 1 African. The SMIC dataset was recorded in an indoor environment designed to resemble an interrogation. Like the procedure of the CASME, participants were demanded to watch highly emotional videos to induce corresponding spontaneous micro-expressions. Each participant was recorded about 50 minutes.

Not all participants showed micro-expressions. The SMIC-HS dataset contains 164 micro-expression video clips elicited from 16 subjects (mean age is 28.1 years, 6 females and 10 males, 8 Caucasians and 8 Asians), while both the SMIC-VIS and SMIC-NIR consist of 71 micro-expression video clips elicited from 8 subjects. For each subject, the number of micro-expression clips ranges between 2 and 29. These expressions are classified into three categories : positive (happiness), negative (sadness, fear and disgust) and surprise. The HS contains 51 positive, 70 negative and 43 surprise. Both the VIS and NIR consist of 28 positive, 23 negative and 20 surprise. Video clips provided in the HS are no more than 50 frames and are within 13 frames both in the VIS and NIR dataset. Clips without micro-expressions randomly selected from the original videos were supplied as the counterpart data for the micro-expression detection. Fig. 2.8 shows samples from the SMIC-HS, in which **(a)** and **(b)** show samples with micro-expressions, in contrast with samples in **(c)** and **(d)**. Faces without micro-expressions can be either neutral or

3. http ://www.oulu.fi/cmvs/node/41319

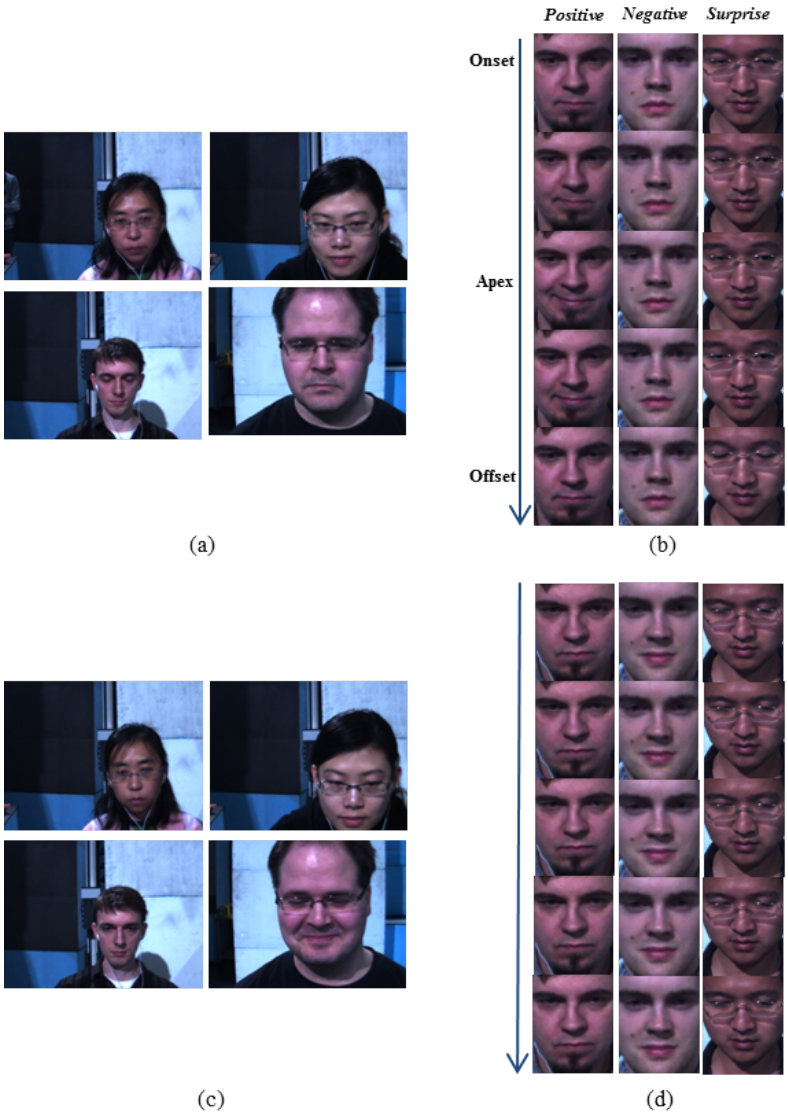


FIGURE 2.8: Samples from the SMIC-HS dataset. Images in (a) and (b) are samples containing micro-expressions, while (c) and (d) provide samples without micro-expressions as a comparison. (a) and (c) Raw images. (b) and (d) Cropped images, for (b), from top to bottom : *positive*, *negative*, *surprise*. We provide five frames of each expression from the onset to offset. Assume the expression sequence has N frames, the index of these five frames are 1, $\frac{N}{4}$, $\frac{N}{2}$, $\frac{3N}{4}$ and N, respectively. Indexes in (d) are same to that in (b).

with macro-expressions, see Fig. 2.8 **(c)** and **(d)**.

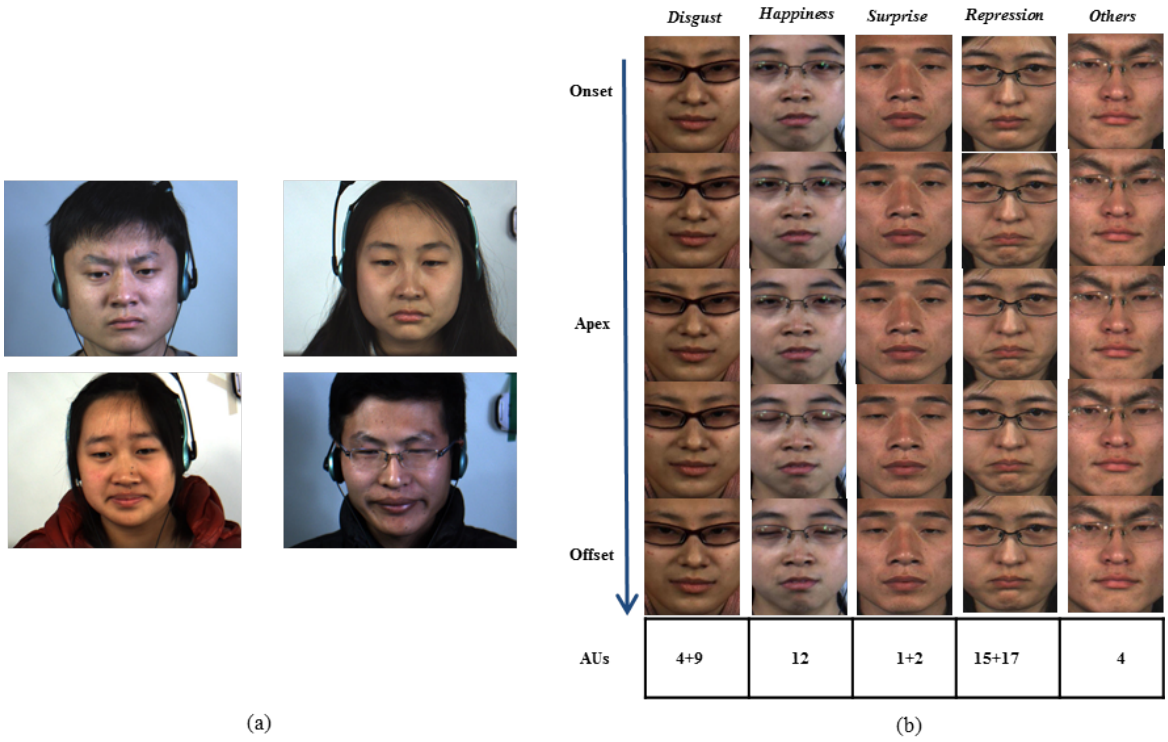


FIGURE 2.9: Samples from the CASME II dataset. (a) Raw images from videos. (b) Cropped images. From left to right : *disgust*, *happiness*, *surprise*, *repression* and *others*. We provide five frames of each expression from the onset to offset labeled with AUs.

CASME II [WJYWZ+14] is an extended version of CASME that offers higher temporal resolution (200fps), larger face size (about 280×340 pixels on facial area), more samples (255 micro-expressions [4]) than the SMIC and CASME datasets. As micro-expression is rapid facial activity with low intensity, the higher spatial and temporal resolution can provide more detailed information on the facial muscle movement. The dataset was built in a well-controlled laboratory environment with a proper illumination that removed light flickering. 35 participants with a mean age of 22.03 years participated to this recording, in which 255 short videos of 26 subjects containing micro-expressions were selected. Each short video involves one micro-expression with the onset, apex and offset frames labeled and action units (AUs) encoded. Five main categories are provided that cover disgust 63, happiness 32, surprise 25, repression 27 and others 99. In fact, there are also two

4. In [WJYWZ+14], the authors reported 247 samples. However, in the excel they provided, we found there 255 samples. In the remaining of the work, we use 255.

small classes of sadness 7 and fear 2 included in the dataset. The 'others' emotion is an ambiguous concept that represents other emotion-related facial movements involving attention or tense. Fig. 2.9 describes some samples from the CASME II dataset.

Chinese Academy of Science Macro- and Micro- expression (CAS(ME)^2) dataset [QWYF16] was established by Xiaolan FU's group from Institute of Psychology, Chinese Academy of Science. In this dataset, researchers used Logitech Pro C920 camera to record 22 participants' (13 females and 9 males) response to nine chosen elicitation videos under two light-emitting diode (LED) lights. The elicitation videos contain two disgust-evoking emotion videos, two anger-evoking emotion videos, and five happiness-evoking emotion video, which range from 1 minute to approximately 2 minutes and 30 seconds. The recorder's resolution was set to 640×480 pixels with 30 frames per second. As the results, 300 macro-expressions with 1303 ms mean duration and 57 micro-expressions with 419 ms mean duration were collected.

Two well-trained FACS coders coded micro-expressions into 28 different AUs with 0.82 coding reliability between each other. These AUs were labeled by four different emotions : positive, negative, surprise and others. The coder manager also provided the onset, apex and offset time of each expression, while arbitrating any disagreement that occurred between the coders.

Spontaneous Micro-Facial Movement (SAMM) [DLC$^+$16] dataset is published by Adrian K. Davison and his colleagues, see a sample set on Fig. 2.11. The SAMM dataset is recorded by a Basler Ace acA2000-340km, with a grey-scale sensor, at 200 fps and 2040×1088 pixels resolution. Two LEDs arrays were applied as the light source to avoid flickering during high speed recording. The majority of the elicitation were video clips chosen by researcher online based on the participant's questionnaire before experiments. Two different duration groups are used to calculate the frequency occurrence for all AUs. In the up to 100 frames group, 222 AUs are detected, while in the up to 166 frames group, only 116 AUs are detected. After the test, three certified coders completed the FACS coding for these videos.

Table 2.1 briefly summarizes these micro-expression databases. For the SMIC, only the second version is listed.

Let us give a review of the procedure of building micro-expressions database. Construction and labeling a good database requests expertise, time and patience. Spontaneous micro-expression data have been collected by recording people's reaction when they watch the film clips with high emotional valence. Researchers provide various clips

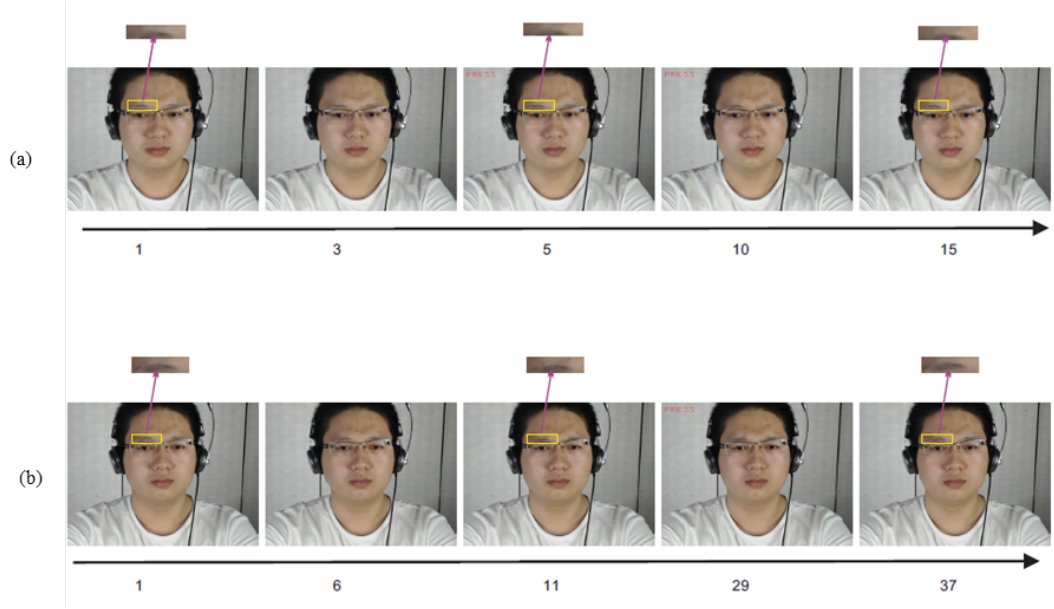


FIGURE 2.10: Examples of micro-expression (a) and macro-expression (b) from the (CAS(ME)^2) dataset. The apex frame appears at about frame 5 for the micro-expression and frame 11 for the macro-expression, which are all negative emotion of anger. The AUs related to these two expressions are all AU 4 (inner brow). Figure is reprinted from [QWYF16].

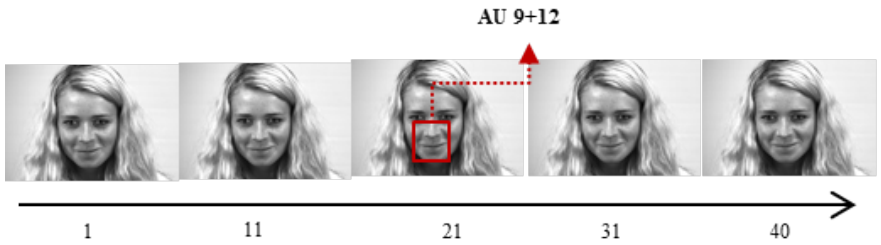


FIGURE 2.11: Examples of micro-expression from the SAMM dataset. The apex frame appears at about frame 11 which represents the positive emotion of smile. The AUs related to the expression is AU $9+12$ (Nose wrinkle and lip corner puller). Figure is reprinted from [QWYF16].

Table 2.1: Summary of micro-expressions databases

Database		#Micro-Expressions	#Participants	Mean Age	#Ethnicities	Fps	Resolution	Elicitation	#Emotion classes	FACS coded
Polikovsky		42	10	\	3	200	640×480	Posed	6	Yes
USD-HD		100	\	\	\	29,7	1280×720	Posed	6	No
York-DDT		18	50	\	\	\	\	Spontaneous	\	No
Canal9		24	195	\	\	\	720×576	Spontaneous	\	No
CASME	A	195	35	22.3	1	60	1280×720	Spontaneous	7	Yes
	B						640×480			
SMIC	SMIC-HS	164	20	26,7	3	100	640×480	Spontaneous	3	No
	SMIC-VIS	71	10			25				
	SMIC-NIR	71	10							
CASME II		255	35	22,03	1	200	640×480	Spontaneous	5	Yes
CAS(ME)^2		57	22	22,59	1	30	640×480	Spontaneous	4	Yes
SAMM		159	32	33,24	13	200	2040×1088	Spontaneous	7	Yes

for producing corresponding micro-expressions. All recordings are produced in laboratory situation with well-controlled lighting condition. Fig. 2.12 shows a straight view for elicitation and recording of micro-expressions. Once the recording is done, it requires to be labeled. Two experienced coders label micro-expressions with the help of subjects themselves (by asking them what emotion they felt). From the databases listed above,

Figure 2.12: Acquisition setup for elicitation and recording of micro-expressions. Figure reprinted from [WJYWZ+14].

it is noted that inducing a wide range of expressions among the subjects is a difficult task. In particular, fear, sadness, and contempt are found to be difficult to elicit in laboratory situation, thus the samples in different categories are distributed unequally. Moreover, the number of categories in each database are different. Fig. 2.13 illustrates

the number of expressions in each category (only listed those categories with many samples that can be used for recognition). Another point to note is the differences among individual subjects. The study of Ekman [Ekm09] indicated that some people might not show micro-expressions, or show few when they are telling lies. Thus, the number of micro-expressions varies across participants. For example, in SMIC dataset, the number of micro-expressions of each participant ranges from 2 to 39.

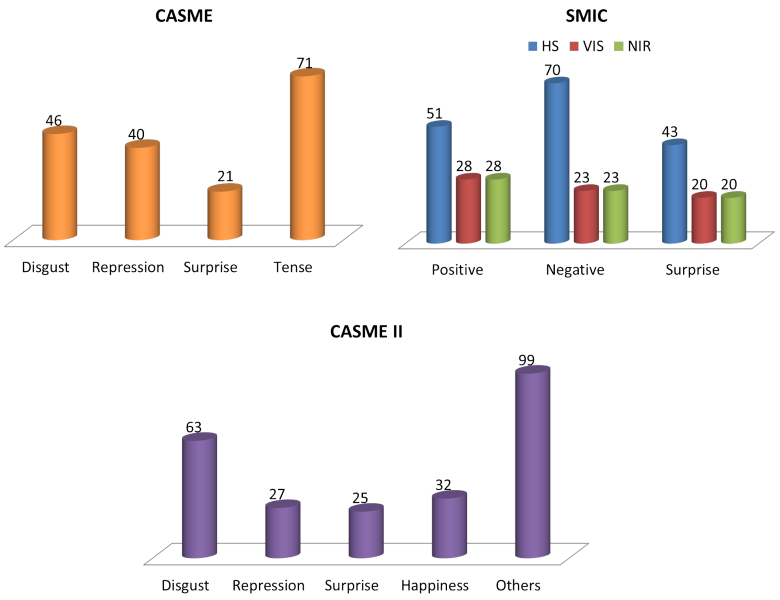


FIGURE 2.13: The number of expressions in each category of the three widely used micro-expressions databases.

A brief note is presented with respect to the available micro-expression databases. The Polikovsky's database contains only posed expressions that may be used for comparison and benchmarking against previous study, it will not be suitable to use it for spontaneous micro-expression recognition. Facial expression data in Canal9 have been collected by recording political debates which contains authentic micro-expressions that can be used for micro-expression detection. The disadvantage of the Canal9 is that this database only provides 24 micro-expressions such that it is insufficient for recognition task. Approaches against illumination changes can be tested in the CASME database. Three datasets including CASME, CASME II and SMIC are popular in micro-expression analysis because they contain sufficient frontal facial expressions labeled by the AUs. With respect to CAS(ME)^2 dataset, it consists of macro- and micro-expressions which can be used for distinguishing micro-expression from macro ones. Diverse ethnicity is

satisfied in SMIC and SAMM, where experiments corresponding to different nations can be conducted.

However, there exists one more unsettled issues. Most databases use young students or teachers who have never criminal experience, restricting the databases to analyze deception in real life, high-stake situation, or medical treatment. It is not possible to find a database for illumination related studies which require various illuminations. Moreover, researches associated to occlusion are important because in real world, partial occlusion appears frequently. The system must be capable of recognizing micro-expressions despite occlusions by sunglasses, facial hairs, hands, scarves, etc. In general, it is difficult to create a database that will satisfy everyone's need. We still look forward new publicly and freely available databases that contain more samples, data under real deception environment, varying conditions of occlusion, lighting, etc. This is important to the future research in this area.

2.6 Conclusion

In this chapter, some basic concepts for the macro- and micro-expression as well as their relations are introduced. The main difference between these two expressions lies at the duration time. The FACS is a powerful tool for the expression analysis. It combines a series of action units to represent various expressions, where the action units are the minimal components for expressions. Most of the existing expression and micro-expression datasets make use of the FACS for coding. In this chapter, we exhaustively introduce nine types of micro-expression datasets, among which the CASME, CASMEII and SMIC datasets which are widely used in micro-expressions analysis and will be used in our experiments. Next chapter will survey the major algorithms that have significantly impacted the development of micro-expression analysis, as well as the powerful image classifier of the Support Vector Machine (SVM).

Chapter 3

Facial feature extraction and classification

3.1 Introduction

Facial feature extraction and classification are the main parts in micro-expression analysis system. Both facial features and machine learning methods are vital for obtaining an excellent performance. Feature extraction is a process of transforming raw data into feature vectors which describes the data properly such that the performance of the model built on the unknown data can be optimal. A feature descriptor is a representation of an image for its characterization. This process involves extracting effective information and ignoring nonessential data. A good feature vector provides essential and discriminative information for tasks like object detection or image recognition. The obtained feature vectors are delivered into image classifiers like Support Vector Machine (SVM) or Random Forest (RF) to produce classification results.

This chapter concentrates both on the feature descriptions and the machine learning classifier. It firstly reviews fundamental descriptors used in micro-expression recognition in Section 3.2, including the local binary patterns based features, the optical flow based features and the histogram of oriented gradient. In Section 3.3 the Support Vector Machine (SVM) classifier for both facial micro-expression detection and recognition is described. The conclusion is presented in Section 3.4.

3.2 Facial Expressions features

This section surveys several basic feature extraction methods which are widely used in micro-expression analysis.

3.2.1 Local Binary Battern (LBP) and its variant

Basic Local Binary Patterns

The local binary pattern was firstly presented by Ojala and Harwood [OPH96], and was proved to be a powerful means of texture description. The operator labels the pixels of an image by thresholding a 3×3 neighborhood of each pixel with the center value and considering the results as a binary number, see an illustration in Fig. 3.1, where the top shows the 8 neighbors around the center pixel g_c, and the bottom presents a detail description of the LBP calculation. Define function

$$s(x) = \begin{cases} 1 & if \quad x \geq 0, \\ 0 & if \quad x < 0, \end{cases} \tag{3.1}$$

where x represents signed differences $(g_p - g_c)$ between neighborhoods g_p and the center pixel g_c.

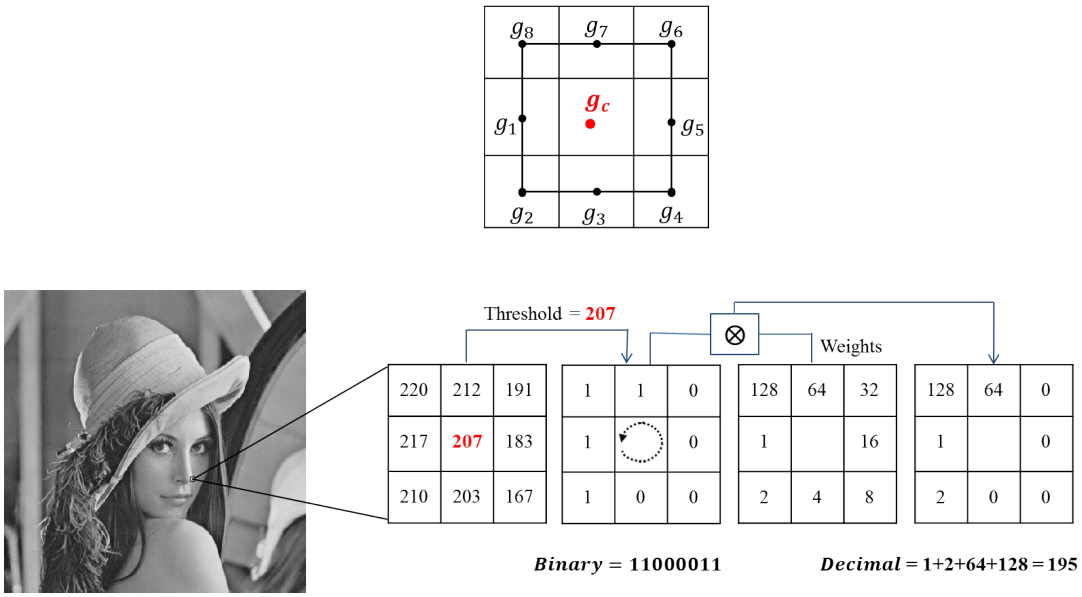


Figure 3.1: An example of the LBP calculation.

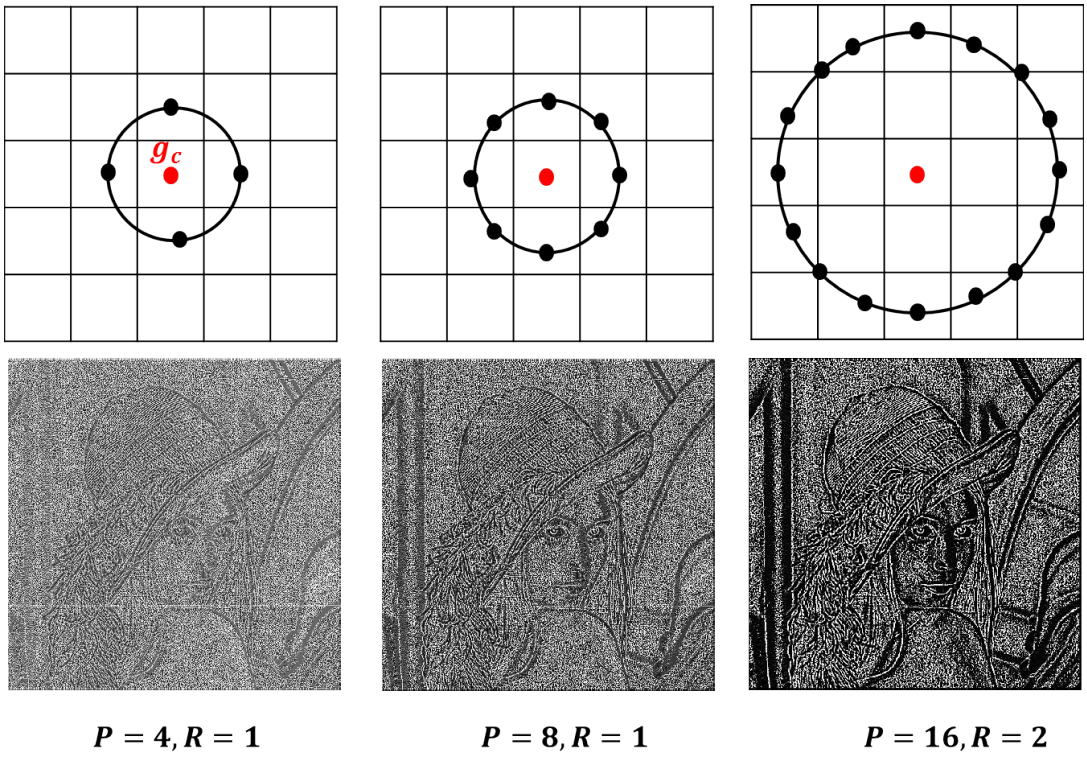


FIGURE 3.2: Different LBP operators. The top describes three circular neighbor opponents for different (P, R). The bottom presents corresponding images of the extracted LBP by applying different (P, R).

Lately, in order to adapt to the texture feature at different scales, the definition of the LBP was extended by Ojala et al. [OPM02], who expanded 3×3 neighborhood to an arbitrary number of neighbors on a circle with a variable radius which is based on double linear differential algorithm. Assuming the central pixel is g_c, a texture model with radius R and sampling number P is constructed, where the number of P and R are variables. In Fig. 3.2, the value of P and R of the LBP operator are ($P = 4$, $R = 1$), ($P = 8$, $R = 1$) and ($P = 16$, $R = 2$), respectively. From the top of this figure, one can notice that there exists the gray values of neighbors which do not fall exactly in the center of pixels at $P = 16, R = 2$, these neighbors are estimated by interpolation.

The local binary number is considered as a micro-texton [HPA04], which describes local texture primitives including spot, flat area, edge, etc (see Fig. 3.3 as an example given $P = 8$).

The theory of circular LBP operator is as follows :

Assuming the coordinates of center pixel is (x_c, y_c), on a circle of radius R, positions

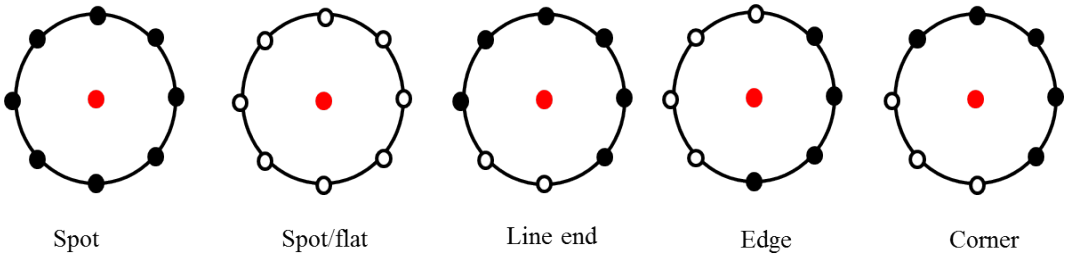


FIGURE 3.3: Examples of texture primitives which can be detected by the LBP (white circles represent ones, black circles zeros, red circles are center points.) [HPA04]

of the P neighborhood points are (X_p, Y_p), $p = 0, 1, \cdots P-1$, then

$$\begin{cases} x_p = x_c - R\cos(\frac{2\pi p}{P}), \\ y_p = y_c - R\sin(\frac{2\pi p}{P}). \end{cases} \tag{3.2}$$

So the local texture feature T of central pixel is defined as a joint distribution

$$T = t(g_c, g_0 - g_c, ..., g_{P-1} - g_c). \tag{3.3}$$

Supposing g_c and g_p are independent of one another, then texture T turns into

$$T \approx t(g_c)t(g_0 - g_c, ..., g_{P-1} - g_c). \tag{3.4}$$

In addition, the distribution $t(g_c)$ in Eq. (3.4) represents the luminance of the entire image, which is unrelated to the local texture description [OPM02]. Thus, the first component can be neglected with respect to luminance of image. This operation makes the LBP operator have strong anti-interference ability against illumination. Hence, texture feature T is represented by the joint difference distribution [OVOP01],

$$T \approx t(g_0 - g_c, ..., g_{P-1} - g_c). \tag{3.5}$$

If only considering the signs of the differences instead of their exact values :

$$T \approx t(s(g_0 - g_c), ..., s(g_{P-1} - g_c)), \tag{3.6}$$

where the $s(x)$ is defined in Eq. (3.1).

Distributing a binomial coefficient 2^P for each sign $s(g_p - g_0)$, then the LBP value of pixel point (x_p, y_p) can be computed by the following formula

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) \times 2^P \tag{3.7}$$

Finally, the feature of an image for texture analysis is represented by a histogram $\hat{\mathcal{H}}$, which can be constructed by

$$\hat{\mathcal{H}}(i) = \sum_{x,y \in \mathbb{Z}^2} I\left\{f(x,y) = i\right\}, \qquad i \in [0, 2^P - 1] \tag{3.8}$$

in which $f(x,y)$ represents the LBP code of center pixel (x,y) and

$$I(x) = \begin{cases} 1 & if \quad x \quad is \quad true, \\ 0 & if \quad x \quad is \quad false. \end{cases} \tag{3.9}$$

Uniform Local Binary Pattern

The LBP operator $LBP_{P,R}$ produces 2^P different output values, corresponding to the 2^P different binary patterns that can be formed by the P pixels in the neighbor set. If all the 2^P patterns are adopted, computation will be very complex. Studies found that some patterns appear in a low frequency ; and some patterns contain more information than others. Therefore, it is possible to use only a subset of the 2^P local binary patterns to describe the texture of images. This types of patterns are called uniform local binary pattern (ULBP) [TTMM00], the formula definition is following :

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=0}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \leq 2 \tag{3.10}$$

From Eq. (3.10), it is noted that the ULBP has a same characteristic if there are at most two changes from 0 to 1 or 1 to 0 in the circular binary code, for example, 00000000 and 11111111 have no binary changes, and 00111100 has two code changes. In fact, only nine set patterns (00000000, 00000001, 00000011, 00000111, 00001111, 00011111, 00111111, 01111111, 11111111) and their circularly rotated versions are uniform patterns. $LBP_{8,1}$ has 256 possible patterns, however, uniform $LBP_{8,1}$ only has 58 possible patterns, which compute 58 bins in computing histogram. Remaining patterns are accumulated into a single bin, which is added to previous 58 bins, resulting into a histogram of 59 bins.

A study conducted in [OPM00] demonstrated that these uniform patterns provide over 90% of texture information.

The number of uniform patterns varies with the number of neighbors P around the center pixel. The ULBP includes in total $(P-1) \times P + 2$ binary patterns, thus resulting into a histogram of $(P-1) \times P + 3$ bins.

Because of its computation speed, the LBP descriptor is widely used in many applications such as texture classification [OPM02], [AHP06, CKM07], facial expression recognition [SGM09] and so on.

So far, the improvement work related to the LBP operator progresses endlessly. Heikkil*ä* et al. [HPS06] proposed the central symmetry local binary pattern (CS-LBP) operator which computes more easily and reduces the data dimension. Tan and his colleagues [TT10] introduced the local ternary patterns (LTP) which is more discriminative and less sensitive to noise than the LBP by changing the two levels qualifications into three levels. Guo et al. [GZZ10] putted forward a complete LBP operator (CLBP) which enhances the LBP operator's description ability of characteristics. Zhao [ZP07] presented a three orthogonal planes (LBP-TOP) operator which allows the LBP to extract texture features from three planes (XY, XT, YT plane).

Local Binary Patterns from Three Orthogonal Planes (LBP-TOP)

The LTP-TOP [ZP07] was proposed to describe dynamic textures. It breaked the limit with which the LBP can only characterize static images and made it possible to apply in space-time analysis, such as dynamic texture classification, dynamic facial expression recognition, etc. The local binary patterns from three orthogonal planes, represent the LBP features extracted from three planes : XY, XT and YT. Fig. 3.4 illustrates example image from three planes. In addition to the XY plane, the XT plane visualizes the changes given a fixed row, while the YT plane shows the changes given a fixed column. The LBP features extracted from three planes are named as XY - LBP, XT - LBP and YT - LBP. The feature of the dynamic sequence is constructed by concatenating features from the three planes into a single histogram.

Recall that the radius R is equal in the process of the LBP feature extraction, which is also appropriate in the LBP computation of the XY plane. However, it is not reasonable for the XT and YT plane, since the number of dynamic sequence frames is usually much less than the resolution of image. Thus, a different radius parameter in space and time is designed, in which the elliptical sampling replaces the traditional circular sampling in XT

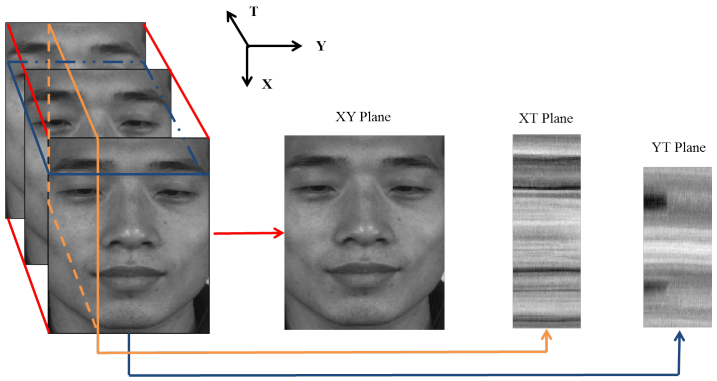


FIGURE 3.4: Example of LBP-TOP three planes [ZP07]. From left to right : A facial expression sequence ; Image in the XY plane (311 × 257) ; Image in the XT plane (311 × 100) in $y = 80$; Image in the YT plane (257 × 100) in $x = 80$.

and YT planes. Suppose the number of neighbors in XY, XT and YT planes is P_{XY}, P_{XT} and P_{YT}, the radii in axes X, Y, and T are R_X, R_Y and R_T, separately, as illustrated in Fig. 3.5. The corresponding feature is denoted as $LBP - TOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$. Given the center pixel $g_{t_c,c}$ and corresponding coordinates (x_c, y_c, t_c), thus, the coordinates of $g_{XY,p}$, $g_{XT,p}$ and $g_{YT,p}$ are given by $(x_c - R_X sin(\frac{2\pi p}{P_{XY}}), y_c + R_X cos(\frac{2\pi p}{P_{XY}}), t_c)$, $(x_c - R_X sin(\frac{2\pi p}{P_{XY}}), y_c, t_c - R_T cos(\frac{2\pi p}{P_{XT}}))$ and $(x_c, y_c - R_Y cos(\frac{2\pi p}{P_{XY}}), t_c - R_T sin(\frac{2\pi p}{P_{YT}}))$, separately.

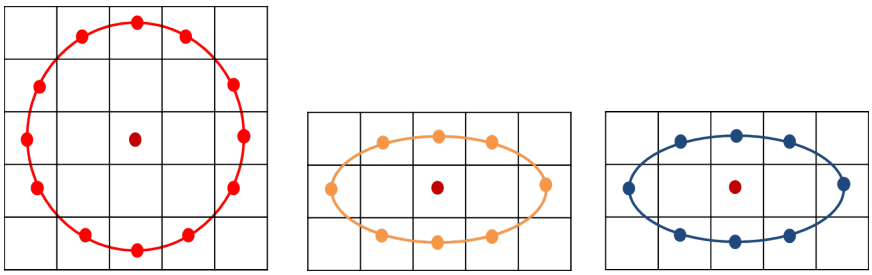


FIGURE 3.5: Different radius parameter sets. From left to right : $R_X = R_Y = 2$ and $P_{XY} = 16$ for XY plane ; $R_X = 2, R_T = 1$ and $P_{XT} = 8$ for XT plane ; $R_Y = 2, R_T = 1$ $P_{YT} = 8$ for YT plane. [ZP07].

A histogram of the LBP-TOP of a dynamic sequence can be denoted by

$$\hat{\mathcal{H}}(i) = \sum_{x,y,t \in \mathbb{Z}^2} I\{f(x,y,t) = i\}, \qquad i \in [0, n_j - 1], j = 0, 1, 2 \tag{3.11}$$

where $(f(x, y, t))$ represents the LBP-TOP code of central pixel (x, y, t), n_j is the number of different labels produced by the LBP operator in the jth plane ($j = 0$ corresponds to the XY plane, $j = 1$ to the XT, $j = 2$ to the YT) and the function $I(x)$ is the same in

Eq. (3.9). The normalized histogram is described as

$$\mathcal{N}_{i,j} = \frac{\hat{\mathcal{H}}_{i,j}}{\sum_{k=0}^{n_j-1} \hat{\mathcal{H}}_{k,j}} \tag{3.12}$$

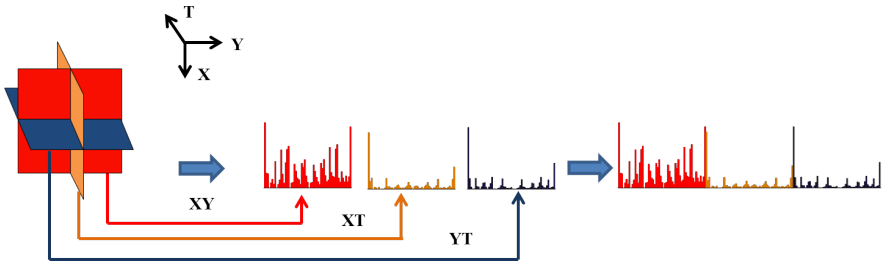


FIGURE 3.6: The process of extracting the LBP-TOP feature. From left to right : three planes of sequence ; the LBP histogram from each plane ; the concatenated LBP-TOP histogram [ZP07].

These three histograms are concatenated to construct a global description of dynamic sequence in order to take into account all spatial and temporal information. This process is illustrated in Fig. 3.6.

Next, another feature of histogram of oriented gradient (HOG) will be introduced, which is also widely used in computer vision area for object detection and recognition.

3.2.2 Histogram of Oriented Gradient (HOG)

The histogram of oriented gradient (HOG) is a feature descriptor to characterize local object appearance and shape by the distribution of local intensity gradients or edge directions. It is proposed by Dalal and Triggs [DT05] and is widely employed in object detection and image recognition. The process is following (see Fig. 3.7) :

(1) **Preprocessing.** Given an arbitrary image G, it is indispensable to resize the image into a fixed ratio of $1:2$ or $1:1$, such as 64×128 or 256×256. Gamma/Color normalization are performed with power law equalization.

(2) **Computing gradients.** Calculating the horizontal and vertical gradients with the kernels of $(-1,0,1)$ or $(-1,0,1)^T$. The magnitude g and direction θ of gradient is attained by the formula of

$$\begin{cases} g &= sqrt(G_x^2 + G_y^2) \\ \theta &= \arctan \frac{G_y}{G_x}, \end{cases} \tag{3.13}$$

where G_x, G_y are the horizontal and vertical gradient, respectively.

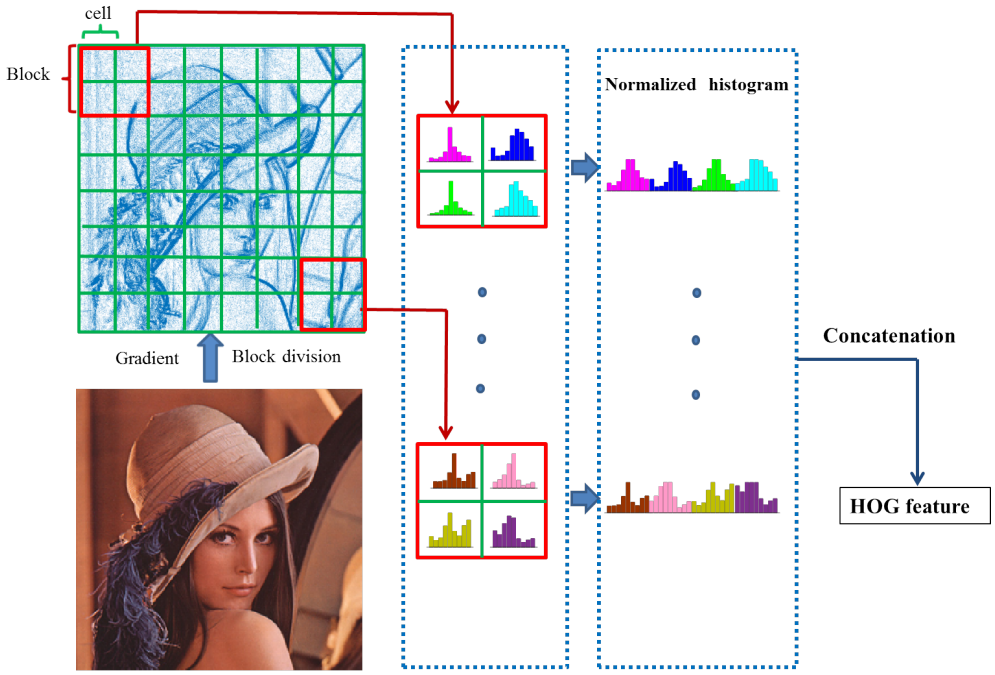


FIGURE 3.7: The process of extracting the HOG feature with the $P = 9$.

(3) **Calculating histogram of oriented gradients.** Dividing the gradient image into small spatial regions ("cells"). Assume the "unsigned gradients" [DT05] are used, where angles are between 0 and 180 degrees with respect to P bins. Four cells form a block, see Fig. 3.7, and for each cell, a local histogram of gradient directions with respect to P bins is accumulated and weighted by its magnitude.

(4) **Normalization.** Performing the normalization on histogram in each block by the L1 or L2 norm.

(5) **Collecting the HOG features over the image.** By concatenating histograms extracted block by block, the HOG descriptor is obtained over the image.

Note that the orientation bins can be evenly spaced over either $0-180$ degrees ("unsigned" gradient) or $0-360$ degrees ("signed" gradient). P can be an arbitrary integer number, as shown in Fig. 3.8, P equals 3, 6 and 9, respectively.

The following subsection presents basic knowledge of the optical flow.

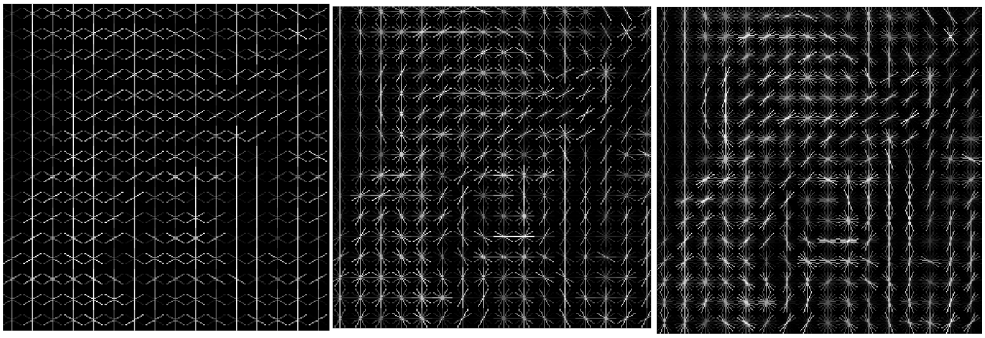


FIGURE 3.8: The HOG feature visualizations with P equals 3, 6 and 9, respectively.

3.2.3 Optical Flow based features

Optical flow

The concept of the optical flow was firstly introduced by Gibson [OGL51] to describe the relation between relative motion of objects and the viewer. For example, you sit on the train and look out the window, an observation is visible that the trees and the buildings are backwards. This movement can be marked as the optical flow. Although the optical flow has been noted since 1950s, Horn and Schunck [HS81] were the first to provide the basic formulations, which have inspired many progresses in the following studies of the optical flow estimation [SRB14]. The optical flow is used to represent the pixel motion's instantaneous of a moving object in the image plane, making use of the changes of images in a sequence over the time domain and the relevance between adjacent frames.

The basic idea of the optical flow is that the intensity value for each point in an object will keep invariant [HS81], which is marked as Brightness Constancy. Let $\Omega \subset \mathbb{R}^2$ be the image domain and $I : \Omega \to \mathbb{R}^+$ be a gray level image. Assume $I(x, y, t)$ is the intensity of a pixel (x, y) at time t, the brightness of a point is constant, so that

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t), \tag{3.14}$$

where $(\delta x, \delta y)$ is the displacement of the local image region at (x, y, t) after time δt. Applying a first-order Taylor expansion to the right-hand side on Eq. (3.14) yields the approximation :

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \tag{3.15}$$

The following notations are introduced

$$I_x = \frac{\partial I}{\partial x}, \quad I_y = \frac{\partial I}{\partial y}, \quad I_t = \frac{\partial I}{\partial t}, p = \frac{dx}{dt}, \quad q = \frac{dy}{dt},$$

where I_x, I_y and I_t as the partial derivations of image intensity with respect to directions x, y and time t respectively. (p, q) is the velocity of a pixel (x, y) at time t.

Based on above Eq. 3.15, the typical formula of the optical flow is expressed as :

$$I_x p + I_y q + I_t = 0, \tag{3.16}$$

Assume $\nabla I = (I_x, I_y)$ is the image gradient field and $\mathcal{F} = (\mathfrak{p}, \mathfrak{q})$ is the optical flow vector field such that the formula can also be written as

$$\langle \nabla I, \mathcal{F} \rangle + I_t = 0, \tag{3.17}$$

which is known as the optical flow constraint equation.

However, it is insufficient to compute two unknown components of u and v by using the optical flow constraint equation. Hence, Horn and Schunck [HS81] provided the additional constraint by minimizing the sum of the squares of the Laplacians of the $\mathfrak{p}$ and $\mathfrak{q}$ components of the flow, which are defined as

$$\begin{cases} \nabla^2 \mathfrak{p} = \dfrac{\partial^2 \mathfrak{p}}{\partial x^2} + \dfrac{\partial^2 \mathfrak{p}}{\partial y^2}, \\ \nabla^2 \mathfrak{q} = \dfrac{\partial^2 \mathfrak{q}}{\partial x^2} + \dfrac{\partial^2 \mathfrak{q}}{\partial y^2}. \end{cases} \tag{3.18}$$

In practice, the problem of solving $\mathfrak{p}$ and $\mathfrak{q}$ becomes to minimize the total error over the image, by combining the Eq. (3.16) and Eq. (3.18) into an objective function,

$$\begin{aligned} E^2 &= w^2 (I_x \mathfrak{p} + I_y \mathfrak{q} + I_t)^2 + (\nabla^2 \mathfrak{p} + \nabla^2 \mathfrak{q}) \\ &\approx (I_x \mathfrak{p} + I_y \mathfrak{q} + I_t)^2 + w^2 [(\overline{\mathfrak{p}} - \mathfrak{p})^2 + (\overline{\mathfrak{q}} - \mathfrak{q})^2] \\ &= E_b^2 + w^2 E_c^2, \end{aligned} \tag{3.19}$$

in which $\overline{\mathfrak{p}}$ and $\overline{\mathfrak{q}}$ are the mean values of $\mathfrak{p}$ and $\mathfrak{q}$, respectively, w is a parameter, and

$$\begin{aligned} E_b^2 &= (I_x \mathfrak{p} + I_y \mathfrak{q} + I_t)^2, \\ E_c^2 &= (\overline{\mathfrak{p}} - \mathfrak{p})^2 + (\overline{\mathfrak{q}} - \mathfrak{q})^2, \end{aligned} \tag{3.20}$$

where the computation of the E_b^2 (data penalty function) is the use of the L2 norm which assumes that the errors in the optical flow constraint equation are Gaussian and

IID [BSL+11]. The E_c^2 is the spatial penalty function which provides one possible solution for computing the image velocity.

The magnitude of the optical flow is defined as

$$M_{\mathcal{F}} = \sqrt{\mathfrak{p}^2 + \mathfrak{q}^2} \tag{3.21}$$

A wide variety of other data and spatial penalty functions have been studied. For data penalty functions, the Charbonnier penalty is a common choice by recent algorithms [BBPW04, WPZ+09], which is a differentiable variant of the L1 norm. Black and Anandan [BA96] introduce the Lorentzian penalty, which is a non-convex robust data penalty. For spatial penalty functions, adding weights to the penalty function is one popular way. The weighting is either isotropic or anisotropic, treating all directions equally or not. Seitz and Back [SB09] present an isotropic penalty function which is down-weighted between different segments. Nagel and Enkelmann [NE86] provide an anisotropic penalty function by adding weight depending on the gradient of the image. Sun et al. [SRB10] design the improved models that weight the neighbors adaptively in an extended image region.

The optical flow involves the movement between two images. Fig. 3.9 illustrates an example of the optical flow computation, where the obvious movements around the mouth can be observed from the second or bottom row.

Histogram of Oriented Optical Flow

Inspired by the success of histograms of features in object recognition, Chaudhry et al. [CRHV09] proposed the histogram of oriented optical flow (HOOF) for the recognition of human actions.

Recall that $\mathcal{F} = (\mathfrak{p}, \mathfrak{q})$ is the optical flow vector field and the corresponding magnitude is $M_{\mathcal{F}} = \sqrt{\mathfrak{p}^2 + \mathfrak{q}^2}$ defined in Equation (3.21).

The orientation space can be either $S^1 = [0, \pi)$ or $S^2 = [0, 2\pi)$. Thus, the angle associated with $\mathfrak{p}$ and $\mathfrak{q}$ which indicates their orientation can be described by a scalar-valued function $\theta_{\mathcal{F}} : \omega \to S^1 = [0, \pi)$

$$\theta_{\mathcal{F}} = \arctan\frac{\mathfrak{q}}{\mathfrak{p}}, \tag{3.22}$$

If the orientation space is S^2, a basic notation is introduced for clarity.

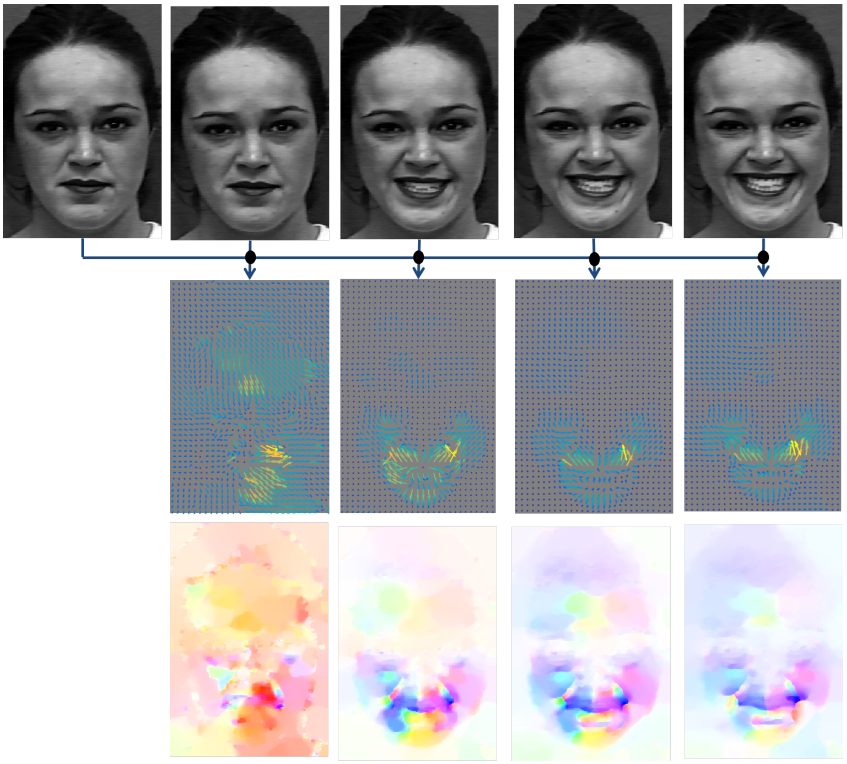


FIGURE 3.9: The optical flow of an image sequence in the facial expression labeled by "Smile" from the extended database (CK+) [LCK+10] by using the algorithm in [SRB10]. Due to the limited space, we only show five frames. The top row represents an image sequence, the second row depicts optical flow field computed from the top row and the bottom row shows the visualization of the optical flow using the color coding scheme in [BSL+11].

Notation We redefine an extended arctan function $\mathfrak{E}_{\arctan}$, which can be mapped to $[0, 2\pi)$ by adding π.

$$\mathfrak{E}_{\arctan}(\beta_2, \beta_1) = \begin{cases} \arctan\left(\frac{\beta_2}{\beta_1}\right), & \text{if } \beta_1 > 0, \\ \arctan\left(\frac{\beta_2}{\beta_1}\right) + \pi, & \text{if } \beta_2 \geq 0, \beta_1 < 0 \\ \arctan\left(\frac{\beta_2}{\beta_1}\right) - \pi, & \text{if } \beta_2 < 0, \beta_1 < 0 \\ 0, & \text{if } \beta_2 = 0, \beta_1 = 0 \\ \pi/2, & \text{if } \beta_2 > 0, \beta_1 = 0 \\ -\pi/2, & \text{if } \beta_2 < 0, \beta_1 = 0 \end{cases} \tag{3.23}$$

Thus, the angle associated with $\mathfrak{p}$ and $\mathfrak{q}$ which indicates their orientation can be described by a scalar-valued function $\theta_{\mathcal{F}} : \omega \to S^2 = [0, 2\pi)$

$$\theta_{\mathcal{F}} = \mathfrak{E}_{\arctan}(\mathfrak{p}, \mathfrak{q}). \tag{3.24}$$

Now we have obtained the orientation function $\theta_{\mathcal{F}}$ and the weighting function $M_{\mathcal{F}}$. The next step is to build the histogram.

Let $\{\Theta_i\}_{1 \leq i \leq P}$ be a collection of connected subsets of the orientation space S^2 satisfying $\Theta_i \cap \Theta_j = \emptyset, \forall i \neq j$ and $\cup_i \Theta_i = S^2$, see Fig. 3.10. Based on such a partition, the

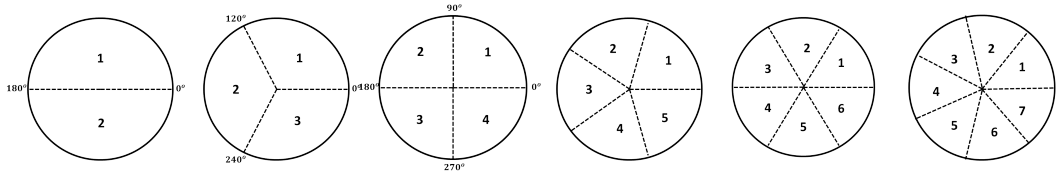


FIGURE 3.10: Example of the $\{\Theta_i\}_{1 \leq i \leq P}$ over the orientation space S^2. The number of P equals 2, 3, 4, 5, 6 and 7 from the left to right.

histogram of optical flow $\mathcal{H}_{\mathcal{F}}$ can be constructed with a set of characteristic functions $\mathcal{B}_i$

$$\mathcal{B}_i(\mathbf{x}) = \begin{cases} 1, & \text{if } \theta_{\mathcal{F}} \in \Theta_i. \\ 0, & \text{otherwise.} \end{cases} \tag{3.25}$$

The weighting function is the norm M such that

$$\mathcal{H}_{\mathcal{F}}(i) = \int_{\Omega} \mathcal{B}_i(\mathbf{x})\, M(\mathbf{x})\, d\mathbf{x}. \tag{3.26}$$

The discrete form $\hat{\mathcal{H}}$ of $\mathcal{H}$ can be expressed by

$$\hat{\mathcal{H}}_{\mathcal{F}}(i) = \sum_{\mathbf{x} \in \mathbb{Z}^2} \mathcal{B}_i(\mathbf{x})\, M(\mathbf{x}), \tag{3.27}$$

for all $\mathbf{x} \in \mathbb{Z}^2$, where $\mathbb{Z}^2$ is the orthogonal discretization grid of the domain Ω.

3.2.4 Video Magnification based features

As the micro-expression is always too small to be caught clearly even through under perfect hardware condition, many researchers use software 'signal amplifier' before recognition. The most popular magnification method in the ME area is Eulerian Video Magnification proposed in [WRS+12], a different direction towards the magnification based on the limited spatio-temporal sensitivity of human naked eyes. The authors exaggerated the subtle color changes of the input video by spatio-temporally process. Meanwhile this process has also demonstrated the ability of magnificating imperceptible motions without any feature tracking or optical flow computation. As their spatio-temporally process was inspired by the Eulerian perspective, this method was named as Eulerian Video Magnification (EVM). The overview of this method was shown as Fig.3.11. Firstly, the input video is decomposed into different spatial frequency bands. Then, a temporal filter is exploited for selecting frequency bands we need. After that, a magnification factor is applied on these filtered bands, which are added back to the original signal and collapsed to generate the output video. It is a technique to reveal subtle changing in videos that are hardly to be observed by naked eyes. Suppose a video sequence is given, the output is an amplified signal to reveal hidden information in video in an indicative manner. This technique can visualize small motion such as flow of bloods and pulse transit more clearly.

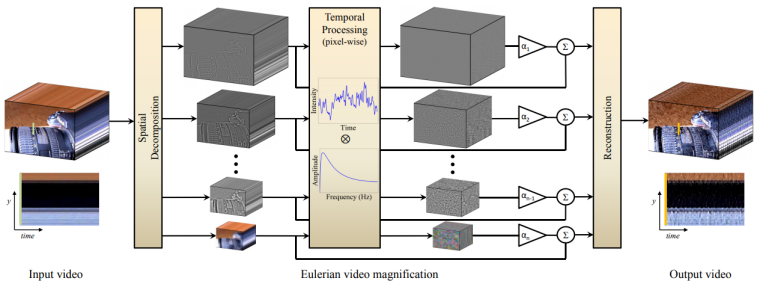


Figure 3.11: Overview of the Eulerian video magnification framework. Figure is reprinted from [WRS+12].

Zarezadeh et.al. [ZR16] studied the potential of this method in the micro-expression area. They use the EVM to retrieve the subtle motions of the face. Their experiments with Spontaneous Micro expression (SMIC) database show that the EVM based method obtained a promising result. Li et al. [LHM+17] applied this method on the motion

magnification of Micro-Expression. The magnification is only applied on recognition task, instead of spotting task, causing that the EVM magnifies unwanted motions such as head movements at the same time. It achieved comparable results by using the EVM based method.

However, the EVM benefits only small magnification factors at high spatial frequencies. Big factors lead to larger scale of motion magnification as well as bigger scale of noise when increasing the magnification factor. So far, only few literatures [ZR16, LHM$^+$17] studied the EVM in micro-expressions analysis because of the influence of noise derived from video magnification.

Next section introduce the support vector machine for micro-expressions classification.

3.3 Facial micro-expression classification using Support Vector Machine (SVM)

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection, which is widely applied in different fields, such as : facial recognition, text and hypertext categorization, image segmentation etc. Till now, the SVM is one of most powerful classifiers for both classification and regression challenges. In our work, the SVM is exploited for conducting both detection and recognition task.

The original SVMs was invented by Vladimir et Vapnik [Vap63], but did not become popular until kernel trick was introduced to it [BGV92], for the kernel trick provides extra nonlinear classification ability to SVMs. And then, Corinna Cortes and Vapnik proposed soft margin to SVMs and expanded its application widely [CV95].

The aim of SVMs is to deal with the data classification problem, so it gives classification between several classes. This is performed by successive binary classifications, which determine the class of a new point based on the data learned from two classes. In specific case of SVMs, a point is treated as a p-dimensional vector, and we want to solve a $(p-1)-$dimensional hyperplane, which is usually called linear classifier, to separate the points.

The main advantages of support vector machines are :

(1) Efficiency in space : it only uses a subset of training points in the decision function.

(2) Efficiency in time for high dimensions problems.

(3) Versatility : it could provide custom kernel for specific problem.

The disadvantages of support vector machines include :

(1) Over-fitting problems when features number is much greater than sample's.

(2) SVMs do not provide direct probability estimates.

Please refer **Appendix A** for detailed description of the SVM.

3.4 Conclusion

A brief literature survey is conducted in this chapter on facial feature extraction and classification. It is well known that feature extraction plays a crucial role in a broad variety of micro-expression analysis applications. Several fundamental methods for feature extraction are reviewed and all of them construct the base of micro-expression feature extraction. Most of the existing works related to the micro-expression analysis involve the improvement or combination of these fundamental methods. A brief description of the SVM is presented. Next Chapter will introduce the micro-expression detection system.

Chapter 4

Micro-Expression Detection

4.1 Introduction

Micro-expression approaches in computer vision area consist of detecting and classifying them in videos. This inspires a series of approaches for micro-expression analysis integrating computer-aided techniques. Most works of the micro-expression analysis concentrate on the classification step [HWZP15, LZY+16], and few works have been devoted to the detection, which is the foundation of this analysis. With the progress of technology, automatic macro-expression detection and recognition can be achieved in real-time and has been successfully applied into business [DCX+15]. Compared to macro-expression, a micro-expression lasts only $40 - 200$ ms, and moreover, its subtle appearance in part of the face makes naked eyes-based detection and recognition difficult to achieve. Given these difficulties suffered by the micro-expression, the facial micro-expression analysis with computer-aider offers a potential solution. The first need is to establish a detection system, in which a robust feature is essential that allows micro-expressions to be discriminated, even in cluttered backgrounds under difficult illumination.

An overview of micro-expressions detection system is summarized in Fig. 4.1.

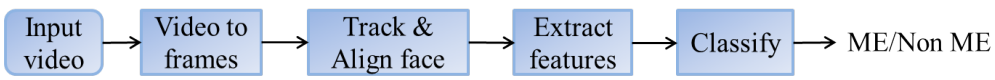


FIGURE 4.1: An overview of micro-expression (ME) detection chain.

The previous researches [PKO09, MZP14a, SBF+14] exploited the HOG, LBP, or OF for spotting micro-expressions in videos. The framework of methods using the HOG and

LBP may not clearly reveal changes in faces, while the framework using the OF can well describe subtle motions in faces but with a high computation cost. Thus, it is possible to develop new feature for micro-expression detection.

This chapter firstly reviews recent related works on micro-expression detection in Section 4.2 . Then, frameworks based on two new features for micro-expression detection are presented in Section 4.3 and 4.4, respectively.

4.2 Related work

So far, several methods have been developed for micro-expressions detection, such as method based on 2D/3D histogram of oriented gradients, local binary patterns and optical flow which will be presented and then used for comparison with the proposed method.

Polikovsky et al. [PKO09] divided the face into different facial regions and used the 3D histogram of oriented gradients descriptor (3D HOG) for feature extraction. The recognition applied the k-means method to cluster the extracted features of each region. The results showed good performance rates (all over 80%) in the regions of the forehead, between the eyes and lower nose. However, the experiments were conducted on a small dataset that only contains 13 posed micro-expressions instead of the spontaneous micro-expressions. Davison et al. [DYL15] used 2D histogram of oriented gradients (HOG) to extract the features of each frame. The chi-squared distance measure was applied to compute dissimilarity between the sequence frames. However, in [DYL15], all detected micro-movements up to 100 frames (200 fps) were classified as true positive including blinks and the eye gaze, without comparing the ground truth of the micro-expression.

Moilanen et al. [MZP14a] adopted local binary patterns (LBP) to extract the features from the blocks of the face. The method relied on calculating the dissimilarity of features for each block by using the chi-squared distance. The detection experiments were conducted on the spontaneous facial micro-expression datasets in order to solve the problem in practice.

Shreve et al. [SBF$^+$14] developed a method for the segmentation of macro- and micro-expression frames by calculating the deformation of facial skin using optical flow (OF). The optical flow is a well-known motion estimation technique and can well spot the subtle movement, but its calculation costs expensive computation time.

Besides, some papers [RHP13, XZW17] addressing the problem of the detection by

training a model to determine if a sequence does or does not contain a micro-expression. Pfister et al. [PLZP11] extracted spatio-temporal local texture features from video sequences and used machine learning algorithms (SVM, MKL, RF) for classification. Ruiz-Hernandez and Pietikäinen [RHP13] encoded the LBP using a re-parametrization of the second local order Gaussian Jet and the SVM for micro-expressions detection and recognition task. Xia et al. [XFP+16] utilized an adaboost model to compute the initial probability for each frame and the correlation between frames in order to generate a random walk (RW) model. The random walk model was used to calculate the deformation correlation between frames and to provide the probability of having micro-expressions in a sequence.

In recent works, Le Ngo et al. [LNSP17] employed the sparse sampling to analyze temporal and spectral structures of spontaneous micro-expressions, and removed neutral faces by the sparse promoting dynamic mode decomposition method (DMDSP).

Instead of developing a training model, we propose a new micro-movement detection method by invoking the integral projection (IP) as a feature descriptor to characterize changes in the blocks of the face. The IP feature is extracted from each individual block which are concatenated. The chi-squared distance is used to measure the IP feature dissimilarity between frames so as to observe for possible micro-expression in the frame sequence. The proposed method is evaluated on two widely used datasets through experimental comparison with some popular feature extractors such as the OF, LBP and HOG. The proposed method is an unsupervised model. One of the main advantages of our model is its low computation complexity : it can obtain comparable or better results than the existing models using OF, LBP and HOG, while requiring much less computation time.

4.3 Micro-expression Detection Using the Integral Projection

4.3.1 Integral Projection

Due to the difficulty for people to read micro-expressions, it is necessary to find appropriate methods for catching subtle and rapid changes of the face. The Integral Projection is presented in the following and it holds as a useful technique for the extraction of facial features. As IP can be extremely effective in determining the position of features, Brunelli

et al. [BP93] applied it for the human face recognition. In a recent work [HZH+16b], a combinational method of the IP and LBP was chosen for micro-expression recognition thanks to its ability for providing the shape property of facial images.

The IP is a simple and rapid feature extraction method which can reduce the 2D image features to a simple 1D data. Let $\Omega \subset \mathbb{R}^2$ be the image domain and $I : \Omega \times D \to \mathbb{R}$ be a sequence of gray level images, where $D \subset \mathbb{R}$ is the time space. At each point $(x, y) \in \Omega$, the intensity value is denoted by $I(x, y)$, and the typical formula of the IP function can be expressed as :

$$IP^H(x) = \frac{1}{y_2 - y_1} \int_{y_1}^{y_2} I(x, y)dy, \tag{4.1}$$

$$IP^V(y) = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} I(x, y)dx, \tag{4.2}$$

where IP_t^H and IP_t^V are the horizontal and vertical integral projection vectors in the rectangle $[x_1, x_2] \times [y_1, y_2]$ at time t, respectively. Fig. 4.2 shows examples of the IP curves (horizontal and vertical). These projections are then concatenated to give the IP feature respectively.

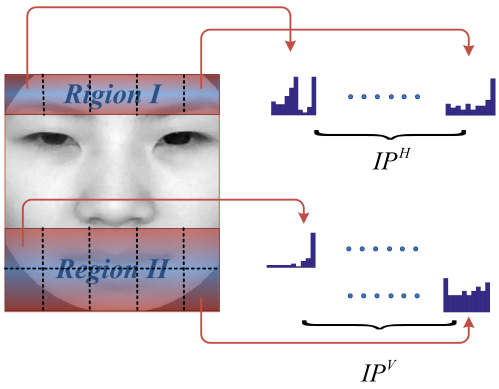


FIGURE 4.2: An example of the IP features. Plots in the first (resp. second) row correspond to the horizontal (resp. vertical) IP function from each block.

4.3.2 Proposed method

The flowchart of the proposed method is summarized in Fig. 4.3 and will be detailed in the following steps. Two main parts are presented in this flowchart : one part is the global pre-processing and featuring, and the other one is the extraction of the micro-expression. The flowchart I includes tracking, registering, cropping, masking, blocking

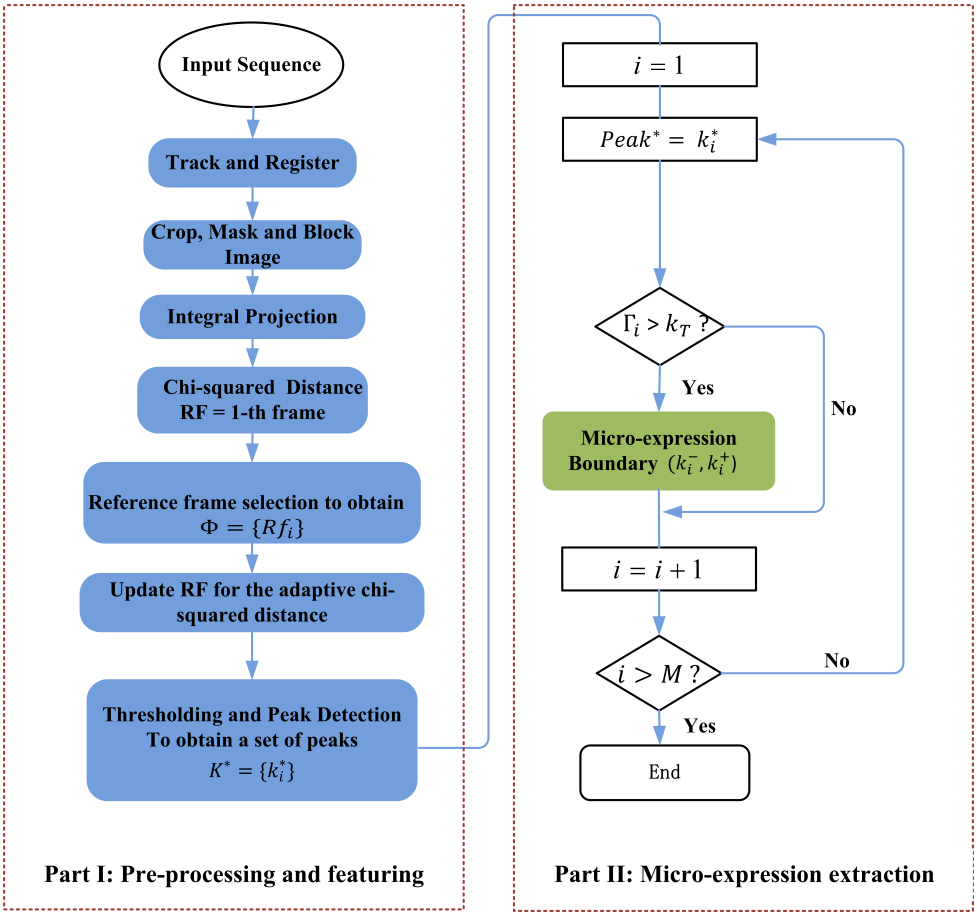


FIGURE 4.3: Flow diagram of the proposed algorithm.

the face, the IP features extraction, chi-squared distance analysis, thresholding, peak detection. The flowchart II consists of the micro-expression extraction.

Face tracking and Processing

Determining the existence and locations of faces in each frame of the sequence is the first step for micro-expression detection. The crucial point of this step is the key points detection for cropping the face. In this section, we choose the Supervised Descent method [XT13] for facial expression points tracking, from which we can obtain 49 facial key points to register and crop the face.

Face alignment operation is necessary for the purpose of keeping eyes in the same line. Since the algorithm depends on careful positioning of the face, alignment step is necessary for the purpose of keeping eyes in horizontal line. By using the facial key points located on the inner eye corners to calculate the angle $\theta \in [0, \pi)$ between the line of the two eyes and a horizontal line, face alignment operation can be performed such that $\theta = 0$. The result of face alignment can be seen in Fig. 4.4b, where the aligned image contains inhomogeneous background, clothes, and hairs, which may influence the detection results. Thus, one can focus on the regions which only contain face information. Let O represent the nose point and d be the distance between E_r and E_l , the central points of the left and right eyes respectively. As illustrated in Fig. 4.4c, we crop a rectangle region K with size $2.2d \times 2d$. The distances from top and bottom boundaries of the rectangle region K to the nose point are $1.3d$ and $0.9d$ respectively. The distances from left and right boundaries of K to the nose point are equal, i.e. both distances are d. The reason of taking nose point as fixed one to cut out the regions of interest is that this point is not easily influenced by subtle facial expressions.

Crop, Mask and Divide face into blocks

The nasal spine point is considered as the fixed point for cropping the face. Face are masked for removing inhomogeneous background and hairs information which may influence the detection results, see Fig. 4.4 (d). During the process of calculating the IP over the whole face, some important spatial information may be missed due to global merging of observations and hence giving difficulties to identify subtle changes of face. Therefore, in order to obtain more accurate spatial information for the detection of micro-expression, two blocked regions of interest (ROIs) are defined for the IP computing :

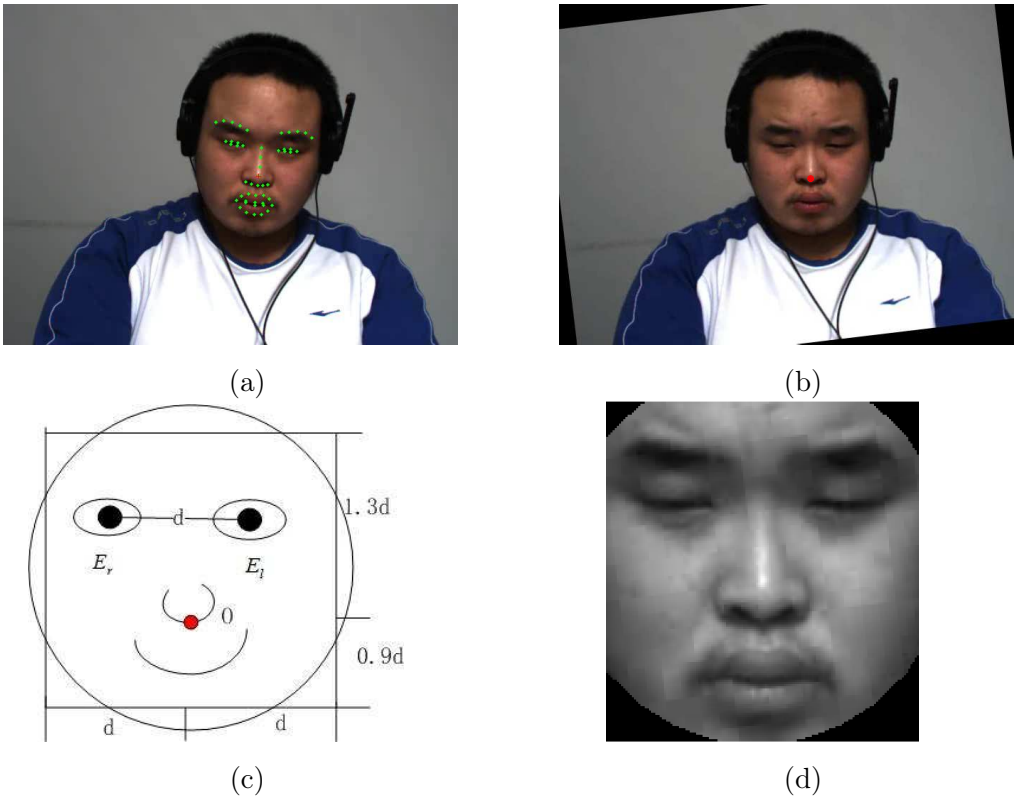


FIGURE 4.4: An example of face detection procedure. **(a)** A detection using Supervised Descent method [XT13]. **(b)** Face alignment. **(c)** Face model for cropping. **(d)** The cropped and masked face.

Region I and Region II respectively involving N and $2N$ blocks, as shown in Fig. 4.2. The number N will be discussed in section 4.3.3. The IP can be calculated in each block to locate small movement for micro-expression analysis.

Feature Extraction Using IP

Once obtained the cropped and blocked face regions, the IP features for each block is computed and then fused together. For the blocks in region I, horizontal IP will describe better the change of the facial skin such as the quick movement of the eyebrow. For the blocks in region II, the micro-movement of the mouth will be well featured by the vertical IP. Thus, the horizontal IP features the region I, while the vertical IP features the region II, as shown in Fig. 4.2.

Feature difference analysis

For feature difference analysis, scanning the sequence, subtraction will be performed between a reference frame (RF) and each successive frame denoted as current frame (CF). This reference frame must be a neutral face or onset frame of a temporal facial expression for highlighting differences along the sequence. Differences will be observed from integral projections. IP_t^H and IP_t^V features are extracted from each frame at each block of the two regions, followed by chi-squared distance computation [MZP14a] to measure the dissimilarity between the CF and the RF frames. The chi-squared distance is an efficient method to compute the distance between the features. Given two IP features of $P = \{p_i\}$ and $Q = \{q_i\}$, the chi-squared distance (CS) [LKR17] is defined by

$$\mathcal{D}_{CS}(P, Q) = \frac{1}{2} \sum_{i=1}^{N} \frac{(p(i) - q(i))^2}{p(i) + q(i)}, \tag{4.3}$$

where N is the length of feature.

The regions I and II generate two chi-squared distance sequences which are denoted by S_1 and S_2, respectively. The computation of $S_j (j = 1, 2)$ for the k-th frame can be expressed as

$$S_j(k) = \mathcal{D}_{CS}(P_0^j, P_k^j) \quad \forall k \in [1, L], \tag{4.4}$$

where P_0^j and P_k^j are the IP features of the very first frame in a video and the k-th frame at the regions I ($j = 1$) and II ($j = 2$). The chi-squared distance S used for micro-movement detection is computed by

$$S(k) = \frac{1}{2}(S_1(k) + S_2(k)), \tag{4.5}$$

which involves the mean values of the normalization of the sequences S_1 and S_2 at the respective location. Normalize the sequences S_1 and S_2 respectively by the values of $\sqrt{\sum_k S_1^2(k)}$ and $\sqrt{\sum_k S_2^2(k)}$.

Reference frames selection

For very long videos segmentation, it is necessary to select different RFs since taking the first frame as the RF will lead to accumulating errors along the sequence. To solve this problem, a new reference frame selection method is proposed. Before the RF selection, one needs to apply low-pass spatial filtering in order to eliminate high frequency details that may influence the result. We give an example in Fig. 4.5a, where we plot the curve (red solid curve) for the chi-squared distance S in Eq. (4.5) when the first frame is selected as the RF. We can see that local maximums of the values of S get larger along the sequence, which may introduce bias in the estimation of the threshold value used for the micro-expression detection.

To cope with this possible bias, we define Φ as a collection of the reference frame indexes which can be expressed as

$$\Phi = \{Rf_i\}_{1 \leq i \leq m}, \quad m \in [1, L-1],$$

where m is the total number of the reference frames and Rf_i is the index of the i-th RF in the sequence. L is the total number of frames in the sequence. Let $Rf_1 = 1$ be the first frame of the sequence then the remaining elements of the collection Φ can be detected in the following two steps.

Firstly, apply the peak detection procedure to the chi-squared distance S in Eq. (4.5) to search for a collection Ψ

$$\Psi = \{\zeta_j\}_{1 \leq j \leq \tau}, \quad \tau \in [1, L],$$

Each element $\zeta_j \in \Psi$ is a local maximizer of the chi-squared distance S. In other words, ζ_j indicates an admissible peak of S such that $S(\zeta_j)$ is a local maximum value which is larger than the mean of S. We further assume that the elements ζ_j of the collection Ψ admits that $\zeta_i < \zeta_j$, if $i < j$. Secondly, search for the nearest local minimum value from the maximum ζ_j along the positive direction. Each pair of adjacent elements ζ_i, $\zeta_{i+1} \in \Psi$ determines a subsequence of frames, among which a local minimizer Rf_i of the computed chi-squared distance S can be obtained. This minimizer is taken as the index of the i-th RF for its notation and it is called Rf_i and $Rf_i \in \Phi$. If there are more than one local

minimizer in the subsequence between ζ_i and ζ_{i+1}, we choose the closest one to the frame ζ_{i+1} (in the sense of Euclidean distance of indexes) as the RF.

Starting from the reference frames collection Φ, a new distance sequence is generated by updating the RF. This is done by computing the chi-squared distance between Rf_i and Rf_{i+1} using Rf_i as RF. Fig. 4.5b is an example of the new chi-squared distance sequence. One can claim that after updating the RF, it is easier to obtain the location of the micro-expression, where the ground truth of the micro-expression given in this example is 39-59 frames. Fig. 4.5c is an illustration of the detected reference frames which are neutral faces or nearly ones.

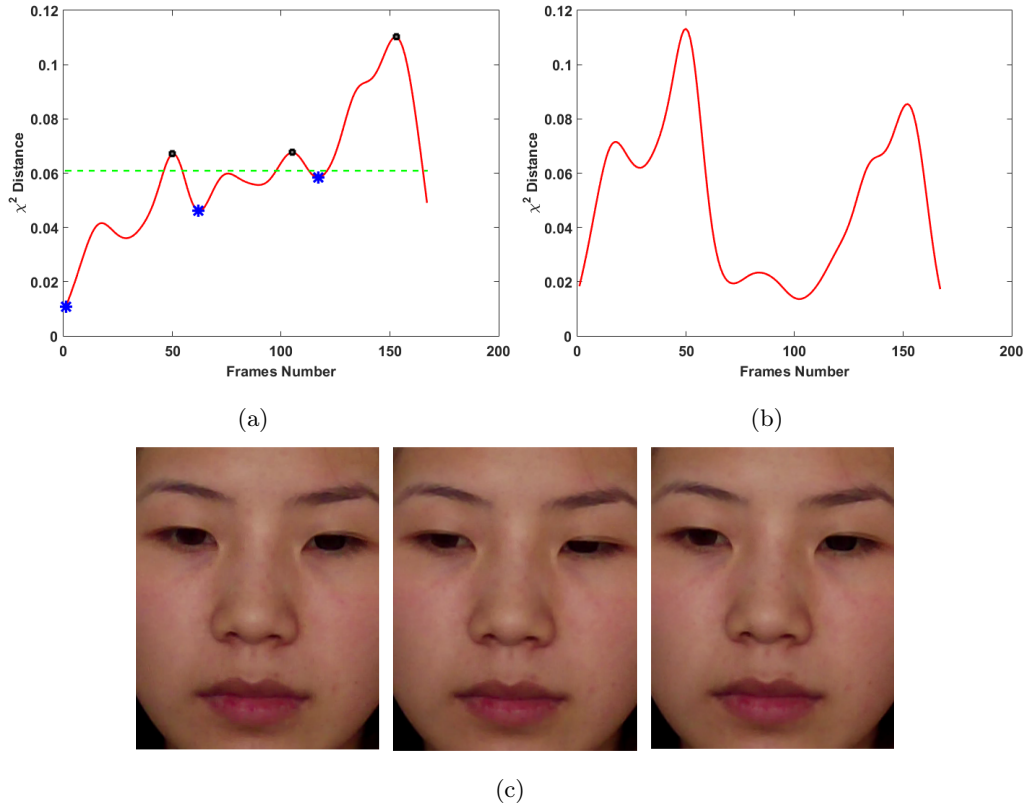


FIGURE 4.5: An example of the detection of Φ. **(a)** The red curve describes the chi-squared distance S. The mean value of S are denoted by a green line. The collection Ψ and Φ are described by black dots and blue star points, respectively. **(b)** illustrates the curve for the chi-squared distance S after updating the RF from the collection Φ. **(c)** shows the three RF of Φ at 1, 62, 117.

In this section, we perform the reference frames selection and the Gaussian smoothing operation on the sequences S to obtain a new adaptive chi-squared distance sequence $S^{'}$ based on the RF collection Φ. The smoothing step aims to remove the noises from the

sequences.

Thresholding and Peak Detection

Once computed the chi-squared distance, it is necessary to use a thresholding method to obtain the location of the micro-expression.

The following steps are applied to the distance sequence $S^{'}$:

(1.) polynomial Fitting. Apply a second order polynomial fitting operation to the sequence $S^{'}$ by the least square method [Shr11] and generate a fitting function ρ. In Fig. 4.6a, we demonstrate the plot curve of the function ρ.

(2.) Micro-expression Appearance Computation. Compute a sequence β with the same length of S by subtracting the fitting sequence ρ from the sequence $S^{'}$. All of the negative values of β are set to 0. The expression of β can be found in Eq. (4.6) and it is illustrated on Fig. 4.6b by using adaptive threshold.

(3.) Peaks Detection. Apply a peak detection procedure to search for a collection of peaks of the sequence β as the indicators of the appearance of micro-expressions. This peaks detection procedure relies on two threshold values as described in the following.

The polynomial fitting step is able to suppress the cropping errors accumulated over the whole sequence. In step 2, the thresholded sequence β is computed as follows :

$$\beta(k) = \max\Big\{S^{'}(k) - \rho(k),\, 0\Big\}, \quad \forall k \in [1, L], \tag{4.6}$$

where ρ is the fitting function and L is the total number of frames in a sequence. The sequence of β involves the information of the existence and the location of the potential micro-expressions.

The peaks detection procedure is carried out dependently of a threshold value T that can be computed by

$$T = \beta_{\text{mean}} + p\,(\beta_{\max} - \beta_{\text{mean}}), \tag{4.7}$$

where β_{mean} and $\beta_{\max}$ are the corresponding mean and maximum values of the thresholded sequence β. The scalar value $p \in [0, 1]$ is a tuning parameter [MZP14b]. This procedure plays the crucial role in the entire course of the micro-expression detection. Thus we give a detailed introduction in the following.

We first detect a collection K^* of M admissible peaks points k_i^* from the sequence β in Eq. (4.6). Each peak point survives in a subsequence $\Gamma_i \subset [1, L]$, where L is the

length of the processed frames including the reference frame. These subsequences Γ_i can be considered as the neighborhoods of the corresponding peak point. We supposed that each subsequence Γ_i has only one peak point and is disjoint to another, i.e.,

$$\Gamma_i \cap \Gamma_j = \emptyset, \quad \forall i \neq j.$$

The detection of the collection K^* and the subsequences Γ_i can be done in two sub-steps.

(1.) First of all, a candidate peak point is a local maximizer of the sequence β within the subsequence Γ_i

$$\beta(k_i^*) \geq \beta(k), \quad \forall k \in \Gamma_i.$$

and has a value of β larger than the threshold T.

(2.) Secondly, we detect the neighborhood Γ_i of this candidate peak point. A subsequence Γ_i can be characterized by the position k_i^* of the candidate peak point and two boundary points k_i^+ and k_i^- such that $\Gamma_i = [k_i^-, k_i^+]$. We search for the position k^+ from the candidate peak point β_i^* along the positive direction till we pass by a point k_* such that $\beta(k_*) > \beta(k_i^*)$, or $\beta(k_*)$ is a local maximum of β, i.e., $\beta(k_*) > \max(\beta(k_* - 1), \beta(k_* + 1))$. Similarly, the position k_i^- is determined along the negative direction.

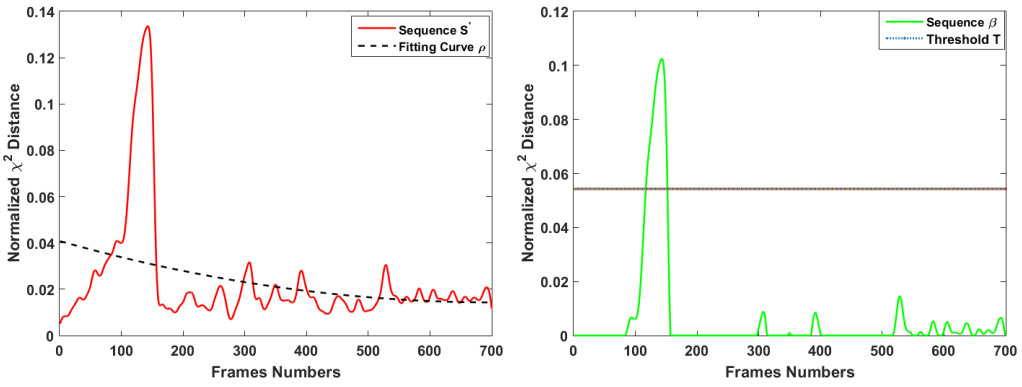
In practice, the value of $\beta(k_*)$ is thought as a local minimum if $|\beta(k_*) - \beta(k_* + 1)|$ is small enough. Based on the two sub-steps described above, a candidate peak point k_i^* is admissible if

$$|k_i^+ - k_i^-| > k_T,$$

where k_T is a given threshold value dependent of datasets. Note that the subsequence Γ_i is actually the i-th duration of micro-movements. The value of $\beta(k_i^*)$ is the i-th value of the peak of the thresholded sequence β. In this step, the values of β at the boundary points k_i^+ and k_i^- are approximately equal to a fraction $\beta(k_i^*)$

$$\beta(k_i^+) \approx \beta(k_i^-) \approx \alpha\, \beta(k_i^*). \tag{4.8}$$

In this paper, the constant α determines the length of detected micro-expression. Fig. 4.6a illustrates, for a video of 700 frames, the fitting curve ρ (dashed line) for $S^{'}$ (red color). The threshold T in Eq. (4.7) and the sequence β in Eq. (4.6) used for spotting the micro-expression are shown in Fig. 4.6b by a horizontal dash line and a green solid curve, respectively. In Fig. 4.6b, it can be seen from the green curve that a micro-expression is



(a) (b)

FIGURE 4.6: An example of the process of the micro-expression detection. **(a)** illustrates the curve of S is fitted by the polynomial fitting and **(b)** provides the step of locating the micro-expression for thresholding the curve of β by T.

spotted around the frame 143. A duration of $128 - 150$ frames is detected with $\alpha = 0.8$ which will be kept inside the algorithm. Compared with the referenced ground truth of frames $131 - 160$ with the peak frame 142, one can see that the detected starting and ending frames are not exactly the same as those of the ground truth but are very close with long overlapping between both. Based on this observation, it is reasonable to claim that obtained results agree with the ground truth.

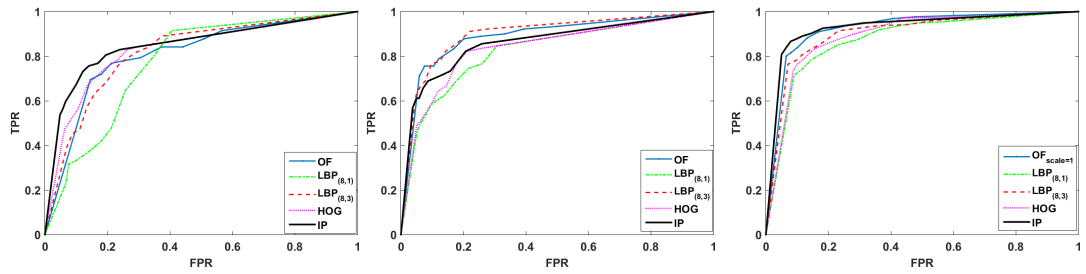
4.3.3 Experiments

For the evaluation, experiments are conducted on two well-known datasets in micro-expression analysis namely CASME [YWL+13a] and CASME II [WJYWZ+14]. Micro-expressions in these two datasets are elicited spontaneously and labeled with reliable ground truth corresponding to the onset, apex and offset frames which can be used for comparisons in the experiments. Please refer to Section 2.5 in **Chapter** 2 for detailed description of CASME and CASME II.

Experiment sets

In our experiments, a comparison with methods of the optical flow (OF), local binary patterns (LBP) and histogram of oriented gradients (HOG) is provided. Parameters are set up in the following.

For the OF, optical flow is computed using the MATLAB implementation of Black



(a) (b) (c)

FIGURE 4.7: **a-c** : ROC curves for the datasets of CASME-A, CASME-B and Casme II, respectively.

[Sun10], [Bla96].

For the LBP, two uniform patterns [Oja96] of $((P,r) = (8,1)$,$(P,r) = (8,3))$ are considered. P corresponds to the number of pixels on the local neighborhood of a circle defined by its radius r.

For the HOG [DYL15], the histogram angle varies from 0 to π or from 0 to 2π, which corresponds to the 'unsigned' or 'signed' gradient. The number of orientation bins is a segmentation value of histogram angles. Here, 8 orientation bins on 2π angle corresponding to signed gradient are chosen as in [DYL15].

The variable parameter $\mathcal{N}$ that defines the number of blocks is set to 5 and α in Eq. (4.8) is set to 0.8. k_T is set to 2 in CASME dataset and set to 7 in CASME II dataset which corresponds to the minimum duration of the micro-expression.

We give a time window tolerance l to detect positively the appearance of the micro-expression peak. The locations of spotted peaks k_i^* are compared with the provided ground truth, and considered to be true positive if they fall within the frame span of $(onset + l, offset - l)$.

We set parameter $l = 5$ for CASME as discussed in [MZP14a], and to $l = 16$ for CASME II same as in [LHM+15]. As eyes are masked in our experiment, spotted eye blinks are counted as false positives not true positives.

Results

Three indicators are adopted for assessing the performance of the algorithm : the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC) and the processing time. The implementation was tested on an Intel Core i7 computer with 16GB of RAM which was equipped with Matlab 2015a.

The ROC curve is used for spotting performance comparison which is illustrated by plotting the true positive rate (TPR) in y-axis against the false positive rate (FPR) in x-axis. The TPR is defined as the number of frames of correctly spotted micro-expression divided by the total number of the ground truth micro-expression frames in the dataset. The FPR is computed as the number of incorrectly spotted frames divided by the total number of non-micro-expression frames in the database. The TPR is in the vertical axis, and the FPR is in the horizontal axis, and p in Eq. (4.7) is used as the varying threshold parameter with step size of 0.1.

Figs. 4.7a to 4.7c show the ROC curves obtained from CASME-A, CASME-B and CASME II for the 4 methods, respectively. Overall, we can observe that the proposed method achieves better performance than other methods (OF, LBP, HOG) in CASME-A and CASME II datasets. Some points with low FPR in ROC curves are meaningful. For example, our proposed method is able to detect 80% of the micro-expression with 4% FPR in CASME II dataset. $\text{LBP}_{(8,3)}$ outperforms $\text{LBP}_{(8,1)}$ on three database and provides the best results in CASME-B.

The area under the ROC curve (AUC) summarizes the spotting performance as shown in Table 4.1. The high values of the AUC means good performance of the method. The AUC results are positive overall and demonstrate that all 4 methods are efficient for spotting micro-expressions. Among the two datasets, a better overall performance can be observed in CASME II. Two reasons can explain this : one is that subjects in CASME dataset often move their head, and another one is that videos are recorded in a different lighting environment in CASME-B leading to the uneven distribution of the lighting in face. In contrast, CASME II contains short video clips at a frame rate of 200fps and no face moving rapidly leading to better detection results.

Among these methods, the proposed method can perform best except in CASME-B dataset because the IP is sensitive to illumination variance while the LBP is robust to illumination. However, the better performance of our method in CASME-A and CASME II shows that the IP is an efficient feature which can describe the temporal dynamic of the micro-expression. The processing time for different methods is presented in the Table 4.2. Here a ratio for computational time comparison is defined as :

$$\gamma = \frac{T_{method}}{T_{IP}}, \tag{4.9}$$

where T_{method} represents the respective processing time for the OF, HOG and LBP. T_{IP} indicates the processing time of the IP features.

Dataset	CASME-A	CASME-B	CASME II
OF	0.8092	0.8888	0.9243
HOG	0.8268	0.8378	0.8939
$LBP_{(8,1)}$	0.7716	0.8244	0.8751
$LBP_{(8,3)}$	0.8177	**0.8987**	0.9014
Proposed IP	**0.8480**	0.8617	**0.9289**

TABLE 4.1: AUC performance for all datasets

Method	Time per frame(ms)	γ
OF	480	631.58
HOG	121.11	159.35
LBP	35.13	46.22
Proposed IP	**0.76**	**1**

TABLE 4.2: Computation time comparison (image size 320×260)

As we can observe in Table 4.2, the algorithm of the optical flow is extremely slow taking 480ms for one image of 320×260 feature extracting. While the proposed method takes only 0.76ms and thus is promising for implementation in real-time process. It is also clear from Table 4.2 that the integral projections provide a huge reduction in computational time. Compared to LBP and HOG, our method still globally outperforms them with a serious advantage in the lower computational complexity.

4.4 Micro-Expression Detection using Facial Geometrical Feature

In this section, a novel method for facial micro-expressions detection is presented. Previous micro-expressions detection features can be categorized into two classes : texture feature (the LBP, HOG, IP) and motion feature (the optical flow), which are extracted on cropped faces that are easily influenced by the alignment operation and cropping

step. Moreover, most of these features have proven highly sensitive to image noise, illumination and pose direction. Inspired by [DCX+15], we propose a new method based on facial geometrical feature for micro-expression detection. It involves an observation of the statistical distance between the geometrical features derived from different video frames. The geometrical feature captures subtle geometric displacements along sequences and is proved to be suitable for different facial analysis tasks that require high computational speed.

The flowchart of the proposed method is summarized in Fig. 4.8. The algorithm initially tracks 49 facial key points in video sequences containing a dynamic micro-expression from the onset till the offset through the apex, using the Supervised Descent method [XT13]. In order to obtain accurate locations, the alignment is performed, please refer to the face tracking and processing part in Section 4.3.2. The main block in the flowchart is "Geometrical distance", which will be explained in details in the following. The displacement analysis is performed for computing difference along the video sequence and the obtained feature is delivered to the SVM for the ME/Non ME classification.

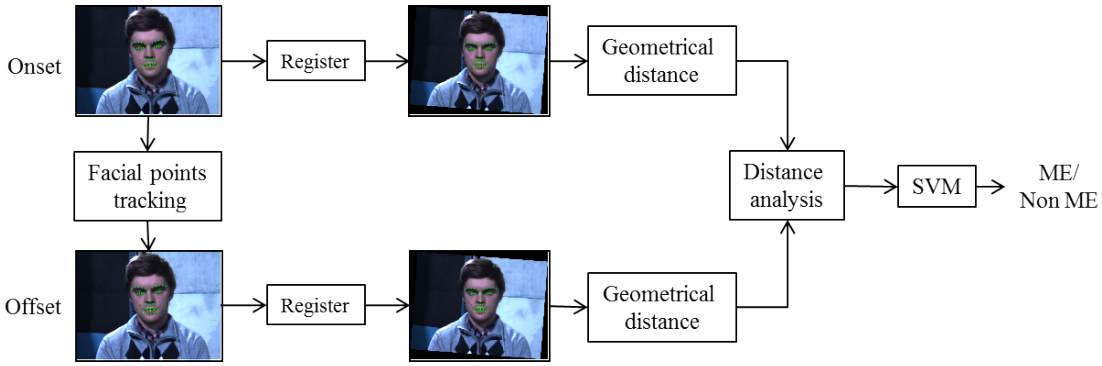


Figure 4.8: Block diagram that summarizes the different steps of the facial geometrical method.

4.4.1 Geometrical Distance

In the proposed method, the classification is performed depending on geometrical displacement information between key points along the video sequence, without taking into account any facial texture information. Among 49 key points, the eyebrow and mouth points are very sensitive to most of the expressions, including 10 points on eyebrows, and 18 points on mouth, see Fig. 4.9. One can think that the geometrical difference of a micro-expression between frames is dynamic along the sequence, it begins from a small

value, then reaches a peak and finally decreases to a small value again. The following will first introduce some basic concepts used in our geometrical feature extraction procedure.

The set of key points of the eyebrow points for the k-th frame are defined as :

$$\mathcal{A}_b^k := \{\mathbf{x}_i^k;\, \mathbf{x}_i^k \in \Omega,\, i = 1, 2, \cdots, n_1\}.$$

where Ω is the image domain. Similarly, for the k-th frame we define the set of key points of the mouth, the upper and lower lip :

$$\begin{aligned}\mathcal{A}_m^k :&= \{\mathbf{x}_i^k;\, \mathbf{x}_i^k \in \Omega,\, i = 1, 2, \cdots, n_2\},\\ \mathcal{A}_{ul}^k :&= \{\mathbf{x}_i^k;\, \mathbf{x}_i^k \in \Omega,\, i = 1, 2, \cdots, n_3\},\\ \mathcal{A}_{ll}^k :&= \{\mathbf{x}_i^k;\, \mathbf{x}_i^k \in \Omega,\, i = 1, 2, \cdots, n_4\},\end{aligned}$$

where $n_1 = \#\mathcal{A}_b^k$, $n_2 = \#\mathcal{A}_m^k$, $n_3 = \#\mathcal{A}_{ul}^k$, $n_4 = \#\mathcal{A}_{ll}^k$ are the cardinal numbers of points involved in collections $\mathcal{A}_b^k$, $\mathcal{A}_m^k$, $\mathcal{A}_{ul}^k$ and $\mathcal{A}_{ll}^k$, satisfying $\mathcal{A}_{ul}^k, \mathcal{A}_{ll}^k \subseteq \mathcal{A}_m^k$. Note that for each frame, we have the same number of key points (resp. 10 eyebrow points, 18 mouth points, 5 upper lip points and 5 lower lip points).

Let $\mathbf{x}_o^k, \mathbf{x}_u^k, \mathbf{x}_{ml}^k, \mathbf{x}_{mr}^k, \mathbf{x}_{lue}^k, \mathbf{x}_{lle}^k, \mathbf{x}_{rue}^k, \mathbf{x}_{rle}^k \in \Omega$ be the reference points which represent the location of nose point, the middle point under the nose, the left and right mouth corner points, the left and right of upper and lower eyelids points in the k-th frame. The set of reference points is summarized as (Fig. 4.9 (a))

$$\mathcal{B}^k = \{\mathbf{x}_o^k, \mathbf{x}_u^k, \mathbf{x}_{ml}^k, \mathbf{x}_{mr}^k, \mathbf{x}_{lue}^k, \mathbf{x}_{lle}^k, \mathbf{x}_{rue}^k, \mathbf{x}_{rle}^k\}, \tag{4.10}$$

We define six distance values :

(1) the distance values between nose point $\mathbf{x}_o$ and the collection $\mathcal{A}_b^k$

$$\mathcal{G}_b(k) := \{\|\mathbf{x} - \mathbf{x}_o^k\|; \mathbf{x} \in \mathcal{A}_b^k\}. \tag{4.11}$$

(2) the distance values between the upper and lower eyelids

$$\mathcal{G}_e(k) := \{\|\mathbf{x}_{lue}^k - \mathbf{x}_{lle}^k\|, \|\mathbf{x}_{rue}^k - \mathbf{x}_{rle}^k\|\}, \tag{4.12}$$

where the $\mathcal{G}_e$ is used for spotting blinks.

(3) the distance values between the middle point under the nose $\mathbf{x}_u$ and $\mathcal{A}_m^k$

$$\mathcal{G}_u(k) := \{\|\mathbf{x} - \mathbf{x}_u^k\|; \mathbf{x} \in \mathcal{A}_m^k\}. \tag{4.13}$$

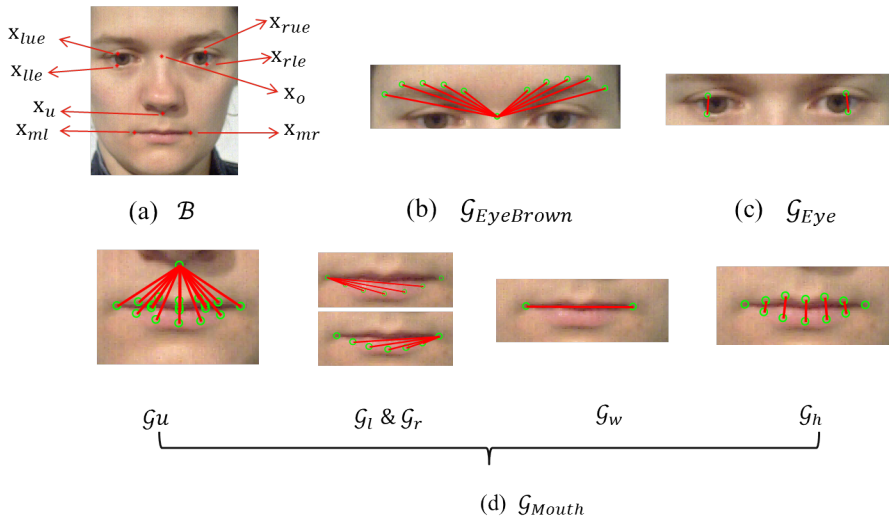


FIGURE 4.9: An illustration of geometrical distances. (a) Locations of reference points. (b) Geometrical distances between eyebrow points and the nose point ($\mathcal{G}_b$). (c) Geometrical distances between eyelids points ($\mathcal{G}_e$). (d) Geometrical distances of the mouth ($\mathcal{G}_m$), including distances between mouth points and the point under nose ($\mathcal{G}_u$), lip corners width($\mathcal{G}_l$ & $\mathcal{G}_r$), mouth width ($\mathcal{G}_w$) and lip height ($\mathcal{G}_h$), respectively.

(4) the distance values between the $\mathcal{A}_{ll}^k$ and $\mathbf{x}_{ml}^k$, the distance values between the $\mathcal{A}_{ll}^k$ and $\mathbf{x}_{mr}^k$

$$\mathcal{G}_l(k) := \{\|\mathbf{x} - \mathbf{x}_{ml}^k\|; \mathbf{x} \in \mathcal{A}_{ul}^k\}, \quad \mathcal{G}_r(k) := \{\|\mathbf{x} - \mathbf{x}_{mr}^k\|; \mathbf{x} \in \mathcal{A}_{ll}^k\} \tag{4.14}$$

(5) the distance value between the $\mathbf{x}_{ml}^k$ and $\mathbf{x}_{mr}^k$

$$\mathcal{G}_w(k) := \|\mathbf{x}_{ml}^k - \mathbf{x}_{mr}^k\|. \tag{4.15}$$

(6) the distance values between the $\mathcal{A}_{ul}^k$ and $\mathcal{A}_{ll}^k$

$$\mathcal{G}_h(k) := \{\|\mathbf{x} - \mathbf{y}\|; \mathbf{x} \in \mathcal{A}_{ul}^k, \mathbf{y} \in \mathcal{A}_{ll}^k\}, \tag{4.16}$$

where $\|\cdot\|$ means the standard ℓ_2 norm. Based on Eq. (4.12), Eq. (4.13), Eq. (4.14) and Eq. 4.15, we define the set of mouth distances in the k-th frame :

$$\mathcal{G}_m(k) = \{\mathcal{G}_u(k), \mathcal{G}_l(k), \mathcal{G}_r(k), \mathcal{G}_w(k), \mathcal{G}_h(k)\}. \tag{4.17}$$

Fig 4.10 shows an example of geometrical distance along the sequence. Within top figure, each red point is derived from the average of all components in Eq. (4.11) and the black point is obtained by calculating the average of all components in Eq. (4.17). One can observe from the figure that the red curve reaches a peak in the 8-th frame which

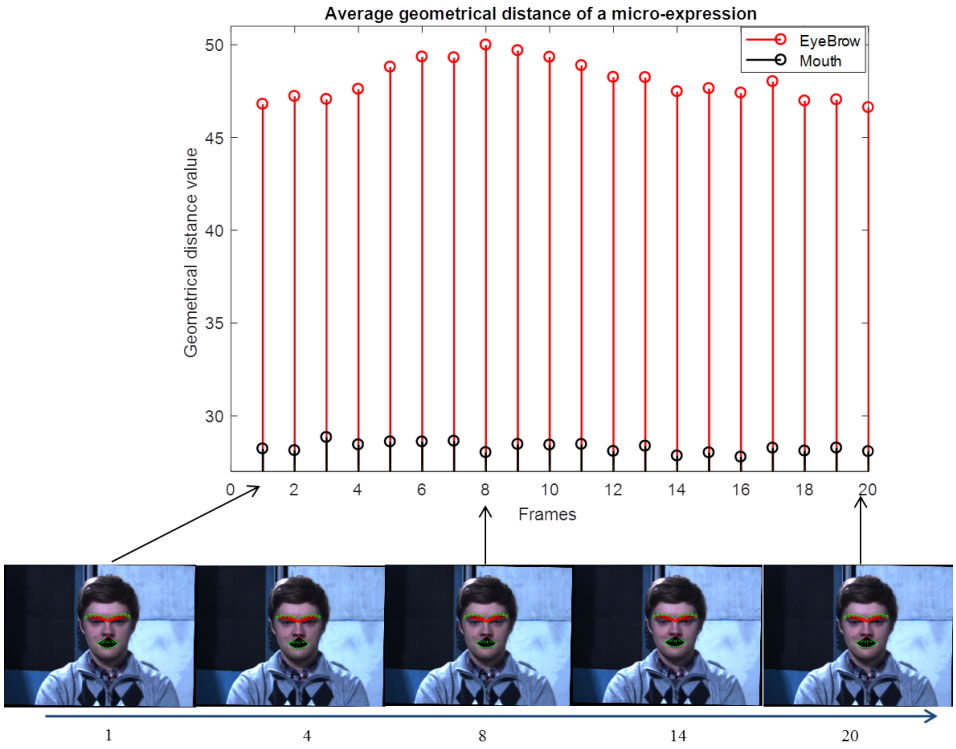


FIGURE 4.10: An illustration of geometrical distance displacement along the sequence. In the top figure, each red point is derived from the average of all components in Eq. (4.11) and the black point is obtained by calculating the average of all components in Eq. (4.17). The bottom represents the original sequence with geometrical distance lines. The sequence is labeled as "surprise"(file $s3\_sur\_02$ of the HS database).

represents the apex of the micro-expression. Then it falls to a lower value at the offset of the micro-expression. But the magnitude difference within the curve is not clearly observed and the curve is not smooth. In order to solve this problem, in the following section, the chi-squared distance as well as the Gaussian smoothing are introduced and performed on geometrical features.

4.4.2 Dissimilarity analysis and Gaussian Smoothing

Recall that the chi-squared distance is performed in Eq. (4.3) for computing dissimilarity of two features of P and Q. Besides, there exists many other measures for the dissimilarity between two features. Except the chi-squared distance, other three distance measures are also investigated, see **Appendix B** for detailed description and results on four datasets. Experiments show that the chi-squared distance outperforms other four distance measures.

In a sequence, the very first frame is selected as the reference frame (RF) and each successive frame denoted as the CF. Differences will be observed from geometrical distance.

The geometrical distance of the eyebrow and mouth generates two distance sequences which are denoted by $\mathcal{G}_b$ and $\mathcal{G}_m$, respectively. The computation of $\mathcal{D}_j(j = b, m)$ for the k-th frame can be expressed as

$$\mathcal{D}_j(k) = \mathcal{D}_{CS}(\mathcal{G}_j(k_1), \mathcal{G}_j(k)) \quad \forall k \in [1, L], \tag{4.18}$$

where $\mathcal{D}_j(k)$ involves the chi-squared distance values between the RF and the k-th frame at the eyebrow region ($j = b$) and mouth part ($j = m$). The $\mathcal{G}_j(k_1)$ and $\mathcal{G}_j(k)$ are geometrical distance of the RF and the k-th frame at the eyebrow region ($j = b$) and mouth part ($j = m$). Since the very first frame is selected as the reference frame, the k_1 is set to 1.

The feature distance $\mathcal{D}$ used for micro-movement detection is computed by

$$\mathcal{D}(k) = max\{\mathcal{D}_b(k), \mathcal{D}_m(k)\}, \tag{4.19}$$

where $\mathcal{D}(k)$ involves the maximum values of the normalization of the sequences $\mathcal{D}_b(k)$ and $\mathcal{D}_m(k)$. The $\mathcal{N}_b$ and $\mathcal{N}_m$ is the normalization of $\mathcal{D}_b$ and $\mathcal{D}_m$ respectively, by applying $\mathcal{N}_b = \frac{1}{n_b}\mathcal{D}_b(k)$ and $\mathcal{N}_m = \frac{1}{n_m}\mathcal{D}_m(k)$, in which $n_b = n_1 = \#\mathcal{A}_b^k$ and $n_m = n_2+n_3+n_4+1 = \#\mathcal{A}_m^k + \#\mathcal{A}_u^k l + \#\mathcal{A}_l^k l + 1$.

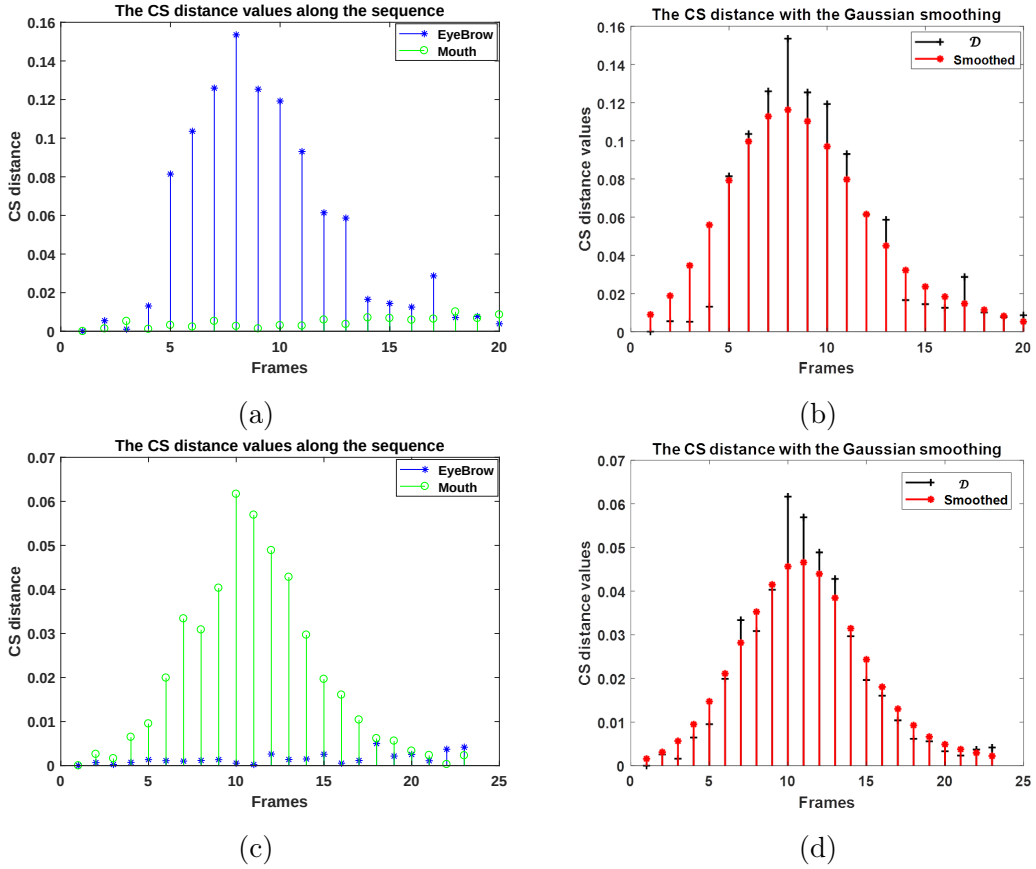


FIGURE 4.11: An illustration of results by applying the chi-squared distance and the Gaussian smoothing operation. In the left figures, the blue curve shows the statistical distances between geometrical features of the eyebrow ($\mathcal{D}_b$) and the green one shows the statistical distances between geometrical features of the mouth ($\mathcal{D}_m$) along the sequence by exploiting the CS. The right figures represent the smoothed curve by applying the Gaussian smoothing on the $\mathcal{D}$. The σ of the Gaussian filter equals to 2.

Generally, the obtained feature distance $\mathcal{D}$ contains some noise needed to be removed. A step can be applied on the $\mathcal{D}$ by convolution with a Gaussian function which is defined as

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{x^2}{2\sigma^2}}, \tag{4.20}$$

where $x \in \mathbb{R}$ and σ is the standard deviation of the distribution.

Fig. 4.11 shows results by applying the chi-squared distance and the Gaussian smoothing operation on the geometrical features of a sequence. The left figures **(a)** and **(c)** show results by exploiting only the chi-squared distance and the right **(b)** and **(d)** show results by applying the Gaussian smoothing. The top figures **(a)** and **(b)** are derived from the sequence labeled as "surprise"(file $s3\_sur\_02$ of the SMIC-HS database). The bottom figures **(c)** and **(d)** are derived from the sequence labeled as "positive"(file $s3\_po\_10$ of the SMIC-HS database). By performing the Gaussian filter on the left curves, the right curves are smoothed which removes noise. The chi-squared distance value between the RF and the CF starts from a small value, then reaches the peak at the 8-th frame, finally falls to a small values again.

4.4.3 Experiments

Experiments using the SVM for micro-expressions detection are conducted upon the SMIC database, which is described in detail in Section 2.5 of the Chapter 2. In fact, there is a previous version of SMIC (referred as SMIC-sub) [1], which only includes 77 video sequences of 6 subjects [PLZP11]. The SMIC [LPH$^+$13] is a full version of the previous data, and contains three subsets, including the SMIC-HS, SMIC-VIS, and SMIC-NIR, which were recorded separately by a high speed (HS) camera of 100 fps, a normal visual camera (VIS) and a near-infrared (NIR) camera at 25 fps, all of three with the resolution of 640 × 480. The HS dataset contains 164 micro-expression video clips elicited from 16 subjects, while both the SMIC-VIS and SMIC-NIR consists of 71 micro-expression video clips elicited from 8 subjects. Clips without micro-expressions randomly selected from the original videos were supplied as the counterpart data for the micro-expression detection.

In our experiments, we use four datasets, including all samples in the SMIC-sub, all micro-expression video clips in the SMIC-HS and SMIC-VIS dataset, 62 video clips of 7 subjects in the SMIC-NIR (all faces of one subject in the video clips are partially

1. http ://www.oulu.fi/cmvs/node/41319

FIGURE 4.12: An example that facial key-points cannot be detected. All faces of this subject in SMIC-NIR dataset are partially appeared.

occulted, as a result, facial key-points cannot be detected, see Fig. 4.12.). The recognition experiments are conducted by applying the leave-one-subject-out (LOSO) where one subject is used as the test set and the others are used as the training set at each loop.

For the SVM, LIBSVM [CL11] is used in our experiments with the radial basis function kernel (RBF)

$$\mathcal{K}_{RBF}(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

and the polynomial kernel (Poly)

$$\mathcal{K}_{Poly}(x_i, x_j) = (\gamma x_i^T x_j + coef)^d$$

, where $\mathbf{x} = \{x_i\}$ is a feature vector, γ, $coef$ and d are parameters in Kernels. To avoid too many combinations of parameters, the fixed values $coef = 2$ and $d = 5$ are adopted based on experiments. In order to find the best parameters, multiples (e.g. 2^k for $k \in [-8, 8]$) of the default value are used as search range in a grid-search using cross-validation to determine C (the penalty parameter) and γ. The highest recognition rate is selected which corresponds to the optimal set of parameters.

We take four selections into account : the RBF kernel, the polynomial kernel, the combination of the RBF and the Gaussian smoothing, the combination of the polynomial and the Gaussian smoothing. For the clarity, only $\sigma = 3$ is exploited. Experiments with respect to $\sigma = 2$ and $\sigma = 4$ will be presented in **Appendix B**.

The number of frames varies depending on the video clips. In order to obtain features with same length, the nearest-neighbor interpolation method is applied to the extracted feature vector. Experiments are conducted on how the recognition rates varies with the

number of features. The number of features varying from 3 to 60 is investigated in our experiment.

Fig. 4.13 to Fig. 4.16 show recognition rates for the SMIC-sub, HS, NIR and VIS dataset, respectively.

Evaluation on SMIC-sub

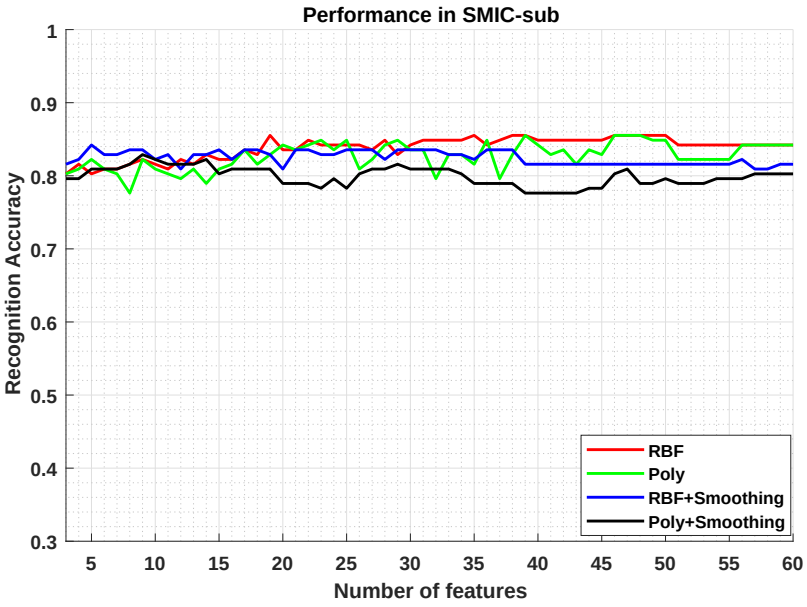


FIGURE 4.13: The results in the SMIC-sub dataset regarding the influence of feature length on the recognition performance.

As observed in Fig. 4.13, the performance of RBF kernel is better than the polynomial kernel in most cases. The combination of the RBF kernel with the Gaussian smoothing operation improves recognition rates at a small number of features, as well as the combination of the polynomial kernel with the Gaussian smoothing. Both the method using the RBF kernel with a number of features 19 (or $46-50$) and the method using the polynomial kernel with a number of features 39 as well as $46-48$ reach the best result 85.53%.

Evaluation on SMIC-HS

Fig. 4.14 shows that the performance of the RBF kernel is still better than the polynomial kernel. A significant improvement by applying the Gaussian smoothing is

observed in this figure. The highest recognition rate 84.15% is achieved by applying the Gaussian filter with the RBF kernel at a number of features 57.

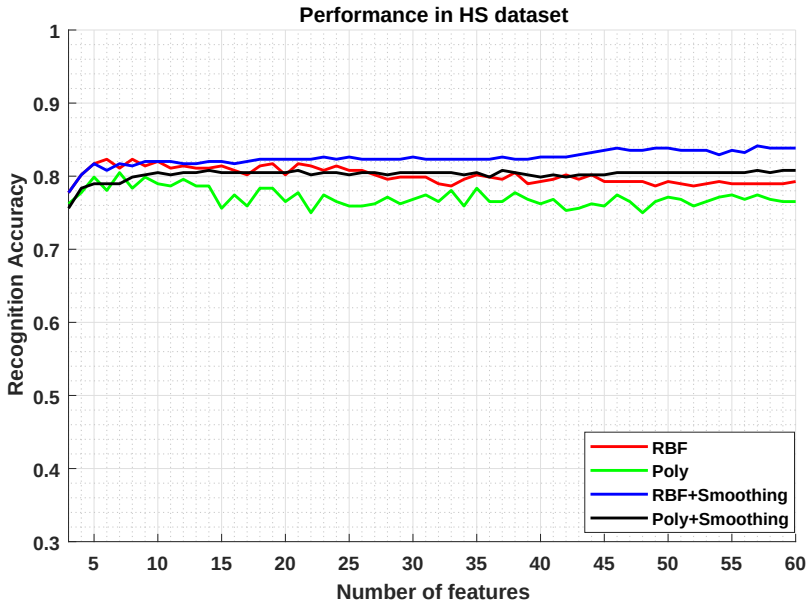


FIGURE 4.14: The results in the SMIC-HS dataset regarding the influence of feature length on the recognition performance.

Evaluation on SMIC-NIR

Fig. 4.15 shows that the RBF outperforms the polynomial kernel in most cases. Almost no improvement by applying the Gaussian smoothing is gained in this figure. The best result by applying only the polynomial kernel is 72.58% at a number of features 10 or 12. The RBF kernel with a number of features 5 as well as 10 yields the highest result 73.39%.

Evaluation on SMIC-VIS

Fig. 4.16 shows the results on the SMIC-VIS dataset regarding the influence of feature length on the recognition performance. The polynomial kernel with a number of feature $23 - 25$ as well as $40 - 42$ achieves the best result 73.24%. The performance of the polynomial kernel is slightly better than that RBF kernel in most cases. Applying the Gaussian filter does not gain any improvements.

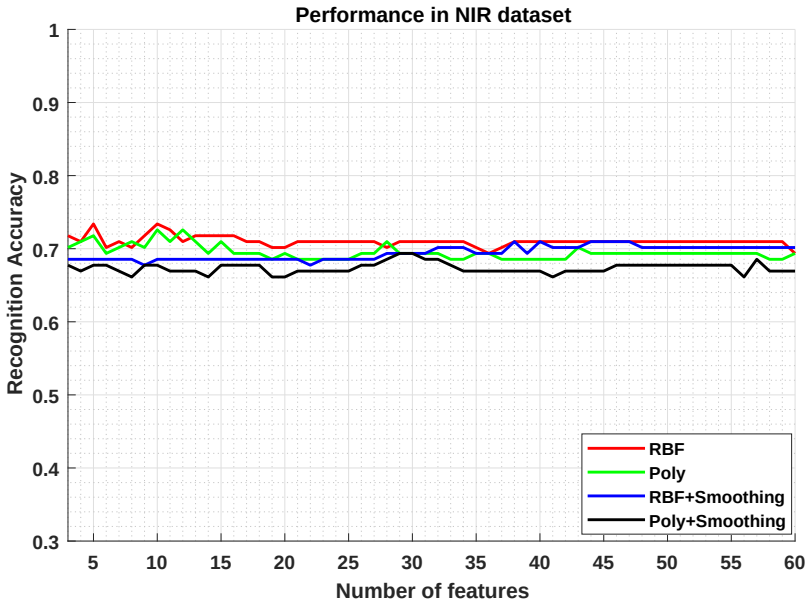


FIGURE 4.15: The results in the SMIC-NIR dataset regarding the influence of feature length on the recognition performance.

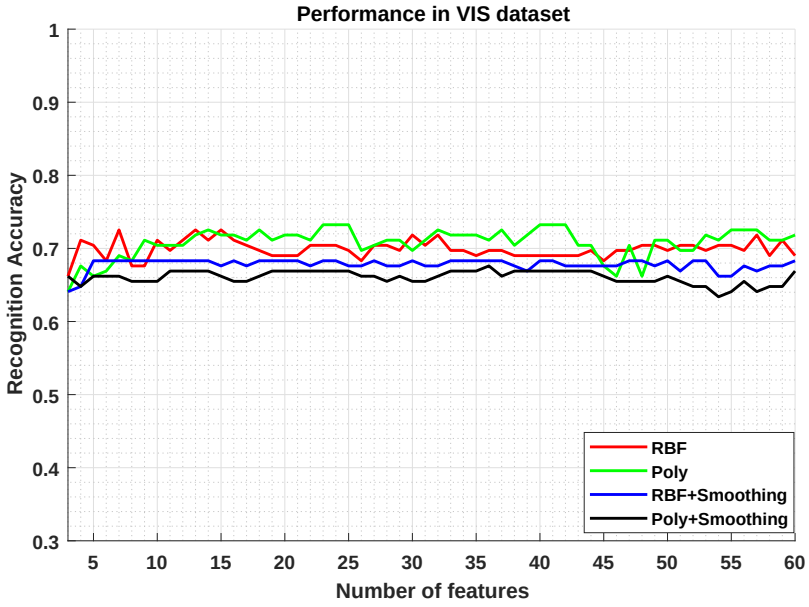


FIGURE 4.16: The results in the SMIC-VIS dataset regarding the influence of feature length on the recognition performance.

Comparison with state-of-the-art

In this section, we compare our results to the state-of-the-art methods (i.e.,TIM [PLZP11], Gaussian jet + LBP [RHP13], SCLQP [HZH+16b] and FDM [XZW17]) in recent literature. Among all compared methods, our method achieves the best recognition rate on all the four datasets. For the NIR dataset, in [XZW17] 8 subjects are used, while our method uses 7 subjects, please see the explanation in Fig. 4.12. From Table 4.3, one can observe that our method is efficient in classifying the ME and the Non ME.

TABLE 4.3: Performance comparison with the state-of-the-art on four datasets. The bold means the highest recognition rate and * means that we directly extract the results from reference papers.

Method	SMIC-sub	SMIC-HS	SMIC-NIR	SMIC-VIS
TIM* [PLZP11]	78,9%	/	/	/
Gaussian jet + LBP* [RHP13]	77,59%	/	/	/
SCLQP* [HZH+16b]	/	75,31%	/	/
FDM* [XZW17]	75,66%	75,3%	72,54%	64,79%
Our method	**85,53%**	**84,15%**	**73.39%**	**73,24%**

4.5 Conclusion

In this chapter, two methods for micro-expression detection are proposed. The analysis on differences of the integral projection allows detecting micro-movements automatically with a low computation complexity. Experimental results are positive on the datasets CASME-A, CASME-B and CASME II, indicating that this method is capable of catching micro-expressions from videos. To our best knowledge, this is the fastest method for automatic micro-expression detection and it could be implemented for future real-time detection. The geometrical feature is extracted from the aligned face. The geometrical distance captures the small relevant motion changes rather than appearance feature. Thus, this feature is robust to illumination variance. The performance on four facial micro-expression datasets (including the SMIC-sub, SMIC-HS, SMIC-NIR, SMIC-VIS) demonstrates the efficiency and discrimination of the geometrical feature. During the experiment, it was noticed that large head motions and the illumination variation can cause mis-detections.

Next chapter will introduce the framework of micro-expressions recognition.

Chapter 5

Motion-based micro-expression recognition

5.1 Introduction

This chapter deals with the problem of automatically recognizing micro-expressions, which has attracted considerable efforts. Effective facial features play a crucial role for micro-expression recognition. These features can be roughly divided into two categories : appearance-based and motion-based methods. Specifically, the appearance-based methods, such as Gabor wavelets [Lee96] and local binary patterns (LBP) [OPH96], are applied to either the whole face or specific face regions to extract the appearance changes of the face. Currently, the main approaches towards the appearance-based methods focus on the use of texture representation [WSF11, DYC$^+$14], due to their computational simplicity. However, it is difficult for the texture features to capture rapid appearance changes reflecting the micro-expression occurring with low intensity. In contrast, the motion features usually derived from the optical flow are considered for micro-expressions recognition [LSPW16]. These methods try to recognize the facial activity by analyzing the motion itself, avoiding referring to the information of static image. Since the micro-expression can be well characterized by the motion features, a better performance compared to appearance-based approaches is observed in [LZY$^+$16]. Another interesting facial features descriptor considering both the appearance- and motion-based information is the LBP-TOP [ZP07].

In this chapter, a new facial feature for micro-expression representation is proposed

for micro-expression recognition. It is known that the optical flow-based features (e.g. histograms of optical flow) require an accurate registration for the recognition applications. However, such requirement is difficult to satisfy in many practical situations. To reduce the influence of inaccurate registration, a new feature descriptor based on the tool of motion boundary is introduced, which is initially proposed for human detection. Motion boundary is calculated by a differential operation on the optical flow vector field which produces two new external vector fields. The proposed feature descriptor is established by fusing the information derived from these external vector fields in a nonlinear mapping manner. For the construction of the respective histograms, we also examine the influence of different ways of building bins in a dense grid. Current studies compute the histograms of motion features by exploiting a fixed number of spatial orientation bins [DT05, WKSL13, LSPW16]. This study investigates the influence of the number of orientation bins and its rotation so as to test the corresponding recognition performance. In order to extract discriminative features, the dimensionality reduction method of the principal component analysis (PCA) is applied since the PCA has powerful properties to identify most meaningful features and maintain a strong correlation between features. Finally, the Support Vector Machine (SVM) is employed for classification.

The rest of the chapter is organized as follows : Section 5.2 briefly reviews the related approaches for micro-expression analysis. In Section 5.3, the proposed optical flow-based features as well as the formal histogram construction method are introduced. The experimental contexts including the computation of the compared descriptors and the parameters setting are presented in Section 5.4. In Section 5.5, the experimental results are discussed. Section 5.6 presents the conclusion.

5.2 Related Works

Micro-expression analysis has attracted much attention from psychologists since the work of Haggard et al. [HI66b] and Ekman et al. [EF69]. It is known that only some experts are capable of observing micro-expressions while most people cannot notice the appearance of these emotions. In order to train people to recognize micro expressions by naked eyes, a micro-expression training tool (METT) was developed by Ekman [Ekm03]. Unfortunately, the study in [EL09b] indicates that good communicators benefited from the METT whereas poor communicators had no gain. These difficulties motivate the approaches of automatic micro-expression analysis [WWQ$^+$17, WYL$^+$15, WYS$^+$16]. For

micro-expression analysis, one of the most important ingredients are the features extracted from the sequence. There are two main categories of approaches for micro-expression recognition.

Local binary patterns inspired approaches. Local binary patterns [OPH96] based features can describe the shape attribute and texture information of face images. The local binary patterns from three orthogonal planes (LBP-TOP) operator was proposed by Zhao et al. [ZP07], where the features are derived from three orthogonal planes. This operator has a strong ability of capturing the dynamic features, leading to a broad variety of LBP-TOP inspired approaches for micro-expression recognition [PLZP11, WYL$^+$14, WYZ$^+$14, DYC$^+$14, LHM$^+$17, JBY$^+$17]. In addition, the improvements of the LBP-TOP operator have been studied in order to deal with different situations. For instances, Wang et al. [WSPO14] proposed a LBP-SIP operator to reduce the redundancy in LBP-TOP patterns. Liet al. [LHM$^+$15] developed an automatic micro-expression analysis system which applied the Eulerian video magnification (EVM) [WRS$^+$12] method to magnify the subtle motions in videos and exploited the dynamic texture features. Huang et al. [HZH$^+$16b] proposed the spatio-temporal completed local spatio-temporal quantized patterns (STCLQP) where the features from three channels containing sign, magnitude and orientation components are extracted and fused depending on discriminative codebooks. Huang et al. [HWZP15] proposed the discriminative spatio-temporal local binary patterns based on an improved integral projection (STLBP-IIP) for micro-expression recognition. This work considered both the shape attribute of face images and the dynamic information of sequences in order to extract more discriminative features. However, it is difficult to capture the subtle appearance changes such as small wrinkles around the eyes or mouths.

Optical flow-based approaches. Using the motion under deformation of faces or facial features is an alternative way to extract facial motion information. Among them, the optical flow-based methods are widely applied because the optical flow provides sufficient dynamic information of a temporal facial expression sequence. Liu et al. [LZY$^+$16] proposed a main direction mean optical flow (MDMO) features for micro-expression recognition. Each face of the sequence is divided into 36 regions of interest (ROIs) partially based on action units and the extracted MDMO features in each region are concatenated into a single feature vector. The MDMO can provide better facial representation and achieve higher recognition accuracy than other existing methods. Liong et al. [LSPW16] exploited the Bi-Weighted Oriented Optical Flow (Bi-WOOF) which emphasize facial motion at

both bin and block levels. In [XZW17], the Facial Dynamics Map (FDM) is proposed to characterize movements of a micro-expression based on the optical flow estimation. Each sequence is divided into spatio-temporal cuboids, and an iterative optimal strategy was developed to calculate the principal optical flow direction in each cuboid for facial expression description.

The above mentioned approaches are non-exhaustive and some recent and relevant works for micro-expression are necessary to be pointed out. In [WSF11, JBY+17], a system was trained based on existing facial expression databases to recognize micro-expressions from micro-expression datasets, providing that a micro and macro-expression have a resemblance appearance. More specifically, Wu et al. [WSF11] developed an automatic micro-expression system by employing facial expression databases as the training set and micro-expression videos from METT [Ekm03] as the test set, respectively. The system utilized Gabor filters to extract facial features and obtained a high accuracy. Wang et al. proposed several methods for micro-expression recognition, including the tensor independent color space (TICS) model in [WYL+14], the combination of the local spatiotemporal directional features (LSTD) and robust PCA (RPCA) in [WYZ+14] and the sparse tensor canonical correlation analysis (STCCA) in [WYS+16]. Jia et al. [JBY+17] proposed a macro-to-micro transformation model which is able to transfer macro-expression learning to micro-expression.

It is still a difficult problem for optical flow to determine the accurate locations of each facial feature mappings between different images even though the face images have been aligned. Such an issue may give rise to wrong orientation and magnitude estimation associated to the optical flow field. In order to address this problem, the motion boundary histograms (MBH) [DTS06] are considered, which is used in action recognition [WKSL13]. Motion boundary is computed by a derivative operation on the optical flow field. It can remove unexpected motions caused by residual mis-registration that appears between images cropped from different frames. Nevertheless, the relative motion can be captured. Based on the the motion boundary, a new descriptor the Fusion Motion Boundary Histograms (FMBH) is introduced.

5.3 Fusion of Motion Boundary Histograms : the proposed descriptor

The facial movement can be well described by associating motion features. In this section, we introduce a new fusion of motion boundary histograms descriptor that is established upon the motion boundary descriptor. Some basic notations are introduced for clarity.

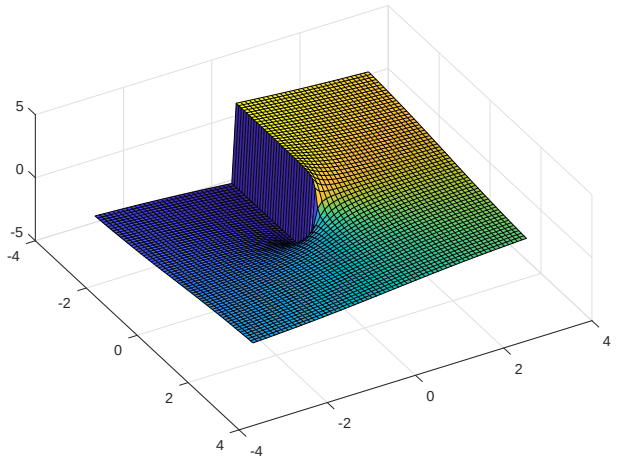


FIGURE 5.1: A 3D graph visualization for the function $\mathfrak{E}_{\arctan}$ defined in Eq. (3.23).

Notation Let $S^2 = [0, 2\pi)$ be the orientation space.

Let $\Omega \subset \mathbb{R}^2$ be an open bounded image domain. Let $\mathcal{F} = (\mathfrak{p}, \mathfrak{q}) : \Omega \to \mathbb{R}^2$ be the optical flow vector field [BA96] with two components $\mathfrak{p}, \mathfrak{q}$ along the x- and y-axis, respectively.

The gradient vector fields of $\mathfrak{p}$ and $\mathfrak{q}$ are respectively denoted by $\nabla\mathfrak{p} = (\mathfrak{p}_x, \mathfrak{p}_y)$ and by $\nabla\mathfrak{q} = (\mathfrak{q}_x, \mathfrak{q}_y)$, where $\mathfrak{p}_x = \partial_x\mathfrak{p}$, $\mathfrak{p}_y = \partial_y\mathfrak{p}$, $\mathfrak{q}_x = \partial_x\mathfrak{q}$, and $\mathfrak{q}_y = \partial_y\mathfrak{q}$.

Our histograms are established over the orientation space $[0, 2\pi)$. The extended arctan function $\mathfrak{E}_{\arctan}$ has been introduced in Eq. (3.23). In Fig. 5.1, we show the 3D graph visualization for the function $\mathfrak{E}_{\arctan}$ defined in Eq. (3.23).

In the following three subsections, we firstly detail the motion boundary features computation and the proposed fusion motion boundary histograms, and then provide the procedure of orientation bins construction.

5.3.1 Motion boundary (MB) features computation

The MB is computed from a differential optical flow which is separated into horizontal and vertical components, as shown in Figure 5.2. The top row shows three key frames of a micro-expression sequence, including onset, apex and offset. The onset is the point at which the expression starts to show up, the apex is the instant when the deformation of the expression reaches the peak and the offset represents the instant when the expression fades away.

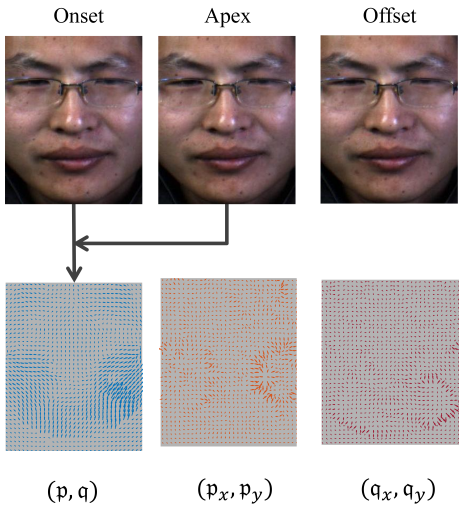


FIGURE 5.2: Illustration of the information captured by optical flow (OF) and motion boundary (MB) descriptor. The top row shows onset, apex and offset frames. The optical flow is computed from the onset and apex frames. The bottom row displays the optical flow vector fields of $(\mathfrak{p}, \mathfrak{q})$, the gradient vector of $\mathfrak{p}$ and $\mathfrak{q}$, respectively. The MB consists of the gradient vector of $\mathfrak{p}$ and $\mathfrak{q}$.

The MBH descriptor [DTS06] is established based on the Euclidean norms and the angle of the vector fields $\nabla\mathfrak{p}$ and $\nabla\mathfrak{q}$. Specifically, the Euclidean norms, respectively denoted by $M_\mathfrak{p}$ and $M_\mathfrak{q}$, of $\nabla\mathfrak{p}$ and $\nabla\mathfrak{q}$, which can be expressed for all $\mathbf{x} \in \Omega$ by

$$M_\mathfrak{p}(\mathbf{x}) = \|\nabla\mathfrak{p}(\mathbf{x})\| = \sqrt{\mathfrak{p}_x^2(\mathbf{x}) + \mathfrak{p}_y^2(\mathbf{x})}, \tag{5.1}$$

$$M_\mathfrak{q}(\mathbf{x}) = \|\nabla\mathfrak{q}(\mathbf{x})\| = \sqrt{\mathfrak{q}_x^2(\mathbf{x}) + \mathfrak{q}_y^2(\mathbf{x})}. \tag{5.2}$$

The angles associated with $\nabla\mathfrak{p}$ and $\nabla\mathfrak{q}$ which indicate their orientation can be respectively characterized by two scalar-valued functions $\theta_\mathfrak{p}, \theta_\mathfrak{q} : \Omega \to S^2$

$$\theta_\mathfrak{p}(\mathbf{x}) = \mathfrak{E}_{\arctan}\left(\mathfrak{p}_y(\mathbf{x}), \mathfrak{p}_x(\mathbf{x})\right), \tag{5.3}$$

$$\theta_\mathfrak{q}(\mathbf{x}) = \mathfrak{E}_{\mathrm{d}\arctan}\left(\mathfrak{q}_y(\mathbf{x}), \mathfrak{q}_x(\mathbf{x})\right). \tag{5.4}$$

The estimation of the histograms of the motion boundary features are discussed in Section 5.4.2.

5.3.2 Fusion motion boundary histograms

The MBHs are calculated in terms of both the norms or magnitudes $M_{\mathfrak{p}}$, $M_{\mathfrak{q}}$ and the angles $\theta_{\mathfrak{p}}$ and $\theta_{\mathfrak{q}}$. We define and consider a scalar-valued function $\alpha : \Omega \to S^2$ which is calculated in terms of $\theta_{\mathfrak{p}}$ and $\theta_{\mathfrak{q}}$ by

$$\alpha(\mathbf{x}) = \mathfrak{E}_{\arctan}(\theta_{\mathfrak{q}}(\mathbf{x}), \theta_{\mathfrak{p}}(\mathbf{x})), \tag{5.5}$$

where $\theta_{\mathfrak{p}}$ and $\theta_{\mathfrak{q}}$ are defined in Eqs. (5.3) and (5.4). The function α can be easily used to establish a new histogram of orientations.

In order to obtain a weighted histogram, the norms $M_{\mathfrak{p}}$ and $M_{\mathfrak{q}}$ are combined together. This can be done by considering a function $M : \Omega \to \mathbb{R}$

$$M(\mathbf{x}) = \sqrt{\mathfrak{p}_x^2(\mathbf{x}) + \mathfrak{p}_y^2(\mathbf{x}) + \mathfrak{q}_x^2(\mathbf{x}) + \mathfrak{q}_y^2(\mathbf{x})} = \sqrt{M_{\mathfrak{p}}^2(\mathbf{x}) + M_{\mathfrak{q}}^2(\mathbf{x})} \tag{5.6}$$

for any $\mathbf{x} \in \Omega$. Actually, the function M in Eq. (5.6) is the Frobenius norm of the Jacobian matrix $\nabla\mathcal{F}$

$$\nabla\mathcal{F}(\mathbf{x}) = \begin{pmatrix} \nabla\mathfrak{p}(\mathbf{x}) \\ \nabla\mathfrak{q}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathfrak{p}_x(\mathbf{x}) & \mathfrak{p}_y(\mathbf{x}) \\ \mathfrak{q}_x(\mathbf{x}) & \mathfrak{q}_y(\mathbf{x}) \end{pmatrix}. \tag{5.7}$$

Now we have obtained the orientation function α and the weighting function M. The next step is to build the new histogram. Since the two histograms used in MBH have been fused together, the proposed method can be named as fusion of motion boundary histograms.

Let $\{\Theta_i\}_{1 \le i \le P}$ be a collection of connected subsets of the orientation space S^2 satisfying that $\Theta_i \cap \Theta_j = \emptyset, \forall i \neq j$ and $\cup_i \Theta_i = S^2$. Based on such a partition, the fusion of motion boundary histogram $\mathcal{H}$ can be constructed with a set of characteristic functions χ_i

$$\chi_i(\mathbf{x}) = \begin{cases} 1, & \text{if } \alpha(\mathbf{x}) \in \Theta_i, \\ 0, & \text{otherwise.} \end{cases} \tag{5.8}$$

and the weighting function which is the norm M such that

$$\mathcal{H}(i) = \int_{\Omega} \chi_i(\mathbf{x})\, M(\mathbf{x})\, d\mathbf{x}. \tag{5.9}$$

The discrete form $\hat{\mathcal{H}}$ of $\mathcal{H}$ can be expressed by

$$\hat{\mathcal{H}}(i) = \sum_{\mathbf{x}\in\mathbb{Z}^2} \chi_i(\mathbf{x})\, M(\mathbf{x}), \tag{5.10}$$

for all $\mathbf{x} \in \mathbb{Z}^2$, where $\mathbb{Z}^2$ is the orthogonal discretization grid of the domain Ω.

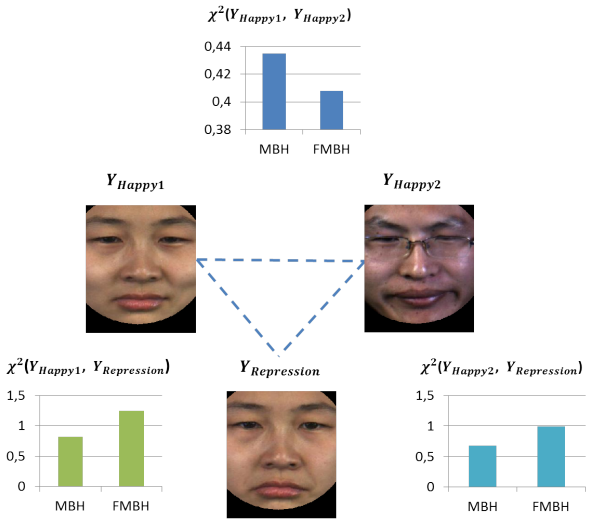


FIGURE 5.3: Illustration of the similarity between facial images by applying the chi-squared distance. The Y_{Happy1}, Y_{Happy2} and $Y_{Repression}$ are from (file path :s9/EP06_02f) (s14/EP09_04) and (s9/EP06_01f) of CASME II, separately.

The dissimilarity measure [1] between two histograms of the features indicating the distance of same expression from two sequences should be small, and the distance is consider to be large for different expression. Here, the chi-squared distance is considered as the dissimilarity measure which can be referred in Eq. (4.3).

An example is presented for illustrating the chi-squared distances comparisons between the MBH and the proposed FMBH in Fig. 5.3. One can point out that the chi-squared distance of the FMBH on Happy expression from two sequences Y_{Happy1} and Y_{Happy2} is lower than that of MBH. While the chi-squared distances of the FMBH between different expressions i.e., $Y_{Repression}$ and Y_{Happy1}, $Y_{Repression}$ and Y_{Happy2} are larger than that of the MBH.

1. Since the feature length of the MBH is twice longer than those resulting from the FMBH, the half values of the chi-squared distances from the MBH are used for fair comparison.

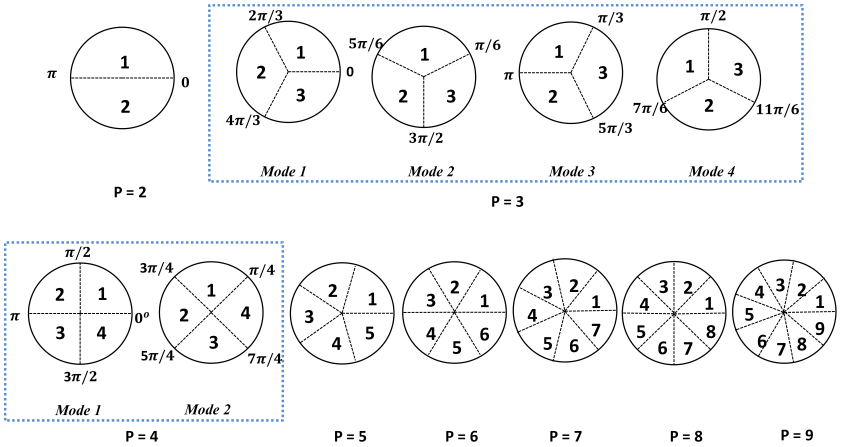


FIGURE 5.4: The bins constructions are from $P = 2$ to $P = 9$ relative to the decomposition of S^2 into $\{\Theta_i\}$ continuous sectors. The $P = 3$ contains four *Modes* and the $P = 4$ consists of 2 *Modes*.

5.3.3 Bins construction

It is known that the performance of feature descriptors relies on the number of bins, as illustrated by Dalal et al. [DT05]. They demonstrated that using 9 bins could generate the best results with respect to the HOG descriptor for human detection. Inspired by this consideration, this section further investigates the influence of both the number of bins and the initial partition angle on the final results. For this purpose, the formal formulation for bins construction are presented in the following.

The disc S^2 is decomposed into a collection of subsets $\{\Theta_i\}$ as described in Fig. 5.4. In practice, it is supposed that these subsets have the same size $|\Theta_i|$, i.e., if given P subsets, $|\Theta_1| = |\Theta_2| = ... = |\Theta_P| = 2\pi/P$. In this case, each subset Θ_i can be constructed by

$$\Theta_i = [\mathcal{A}_{ini} + (i-1)\mathcal{A}_{res},\ \mathcal{A}_{ini} + i\mathcal{A}_{res}), \quad 1 \leq i \leq P, \tag{5.11}$$

where $\mathcal{A}_{ini}$ is the initial partition angle and $\mathcal{A}_{res} = 2\pi/P$ is the size of each subset. Note that $\mathcal{A}_{ini}$ is also the lower bound of the first subset $\Theta_1 = [\mathcal{A}_{ini}, \mathcal{A}_{res})$. From Eq. 5.11, we can see that the subsets $\{\Theta_i\}$ can be determined by a pair $(\mathcal{A}_{ini}, P)$ indicating the partition mode of a disk. In Fig. 5.4, partition modes for different values of $(\mathcal{A}_{ini}, P)$ are displayed. In this section, for $P = 3$, four modes are considered, each of which is constructed respectively by assigning the values of 0, $\pi/6$, $\pi/3$ and $\pi/2$ to the initial angle $\mathcal{A}_{ini}$. While for $P = 4$, we consider two Modes, where $\mathcal{A}_{ini} = 0$ or $\pi/4$.

5.4 Implementation Details

This section first briefly describes micro-expressions datasets that are used in experiments and the experimental protocols conducted upon these datasets. Then, the baseline methods for motion features extraction are introduced.

5.4.1 Datasets

This section brief surveys four micro-expression datasets that are widely used in micro-expression analysis : (i) the Chinese Academy Of Sciences Micro-expression (CASME) [YWL+13a] (ii) the improved CASME (CASME II) [WJYWZ+14] (iii) the Spontaneous Micro-expression Database (SMIC) [LPH+13] and (iv) the Chinese Academy of Sciences Macro-Expressions and Micro-Expressions (CAS(ME)^2). Please refer Section 2.5 in **Chapter** 2 for detailed description of these four databases.

In our experiments, for the CASME, 150 samples from 19 subjects were selected, categorized into four classes : including repression (29 samples), disgust (40 samples), surprise (18 samples) and tense (63 samples). For the CASME II, 246 micro-expressions from 26 subjects in CASME II dataset were used, categorized into five classes : disgust (63 samples), happy (32 samples), repression (27 samples), surprise (25 samples) and others (99 samples). For the SMIC, we conduct experiments on the HS subset which includes 51 positive, 70 negative and 43 surprise samples. For the CAS(ME)^2, all 357 video clips are used.

Figure 5.5 illustrates some micro-expressions from the four datasets. It is noted that the changes among onset, apex and offset are hardly noticeable for human eyes.

Our motion features are extracted between two frames, the onset and the apex frame. Note that the original paper [LPH+13] did not provide the ground truth index of apex frame in SMIC database, the procedure for determining the index of apex frame is presented in the algorithm 1. The Fig. 5.6 is the illustration of the step 7 and step 8 in the algorithm 1.

A Micro-Expression Apex Detection Procedure

Since the original paper did not provide the ground-truth index of apex frame in SMIC database, the procedure used for determining the index of apex frame is presented in the following.

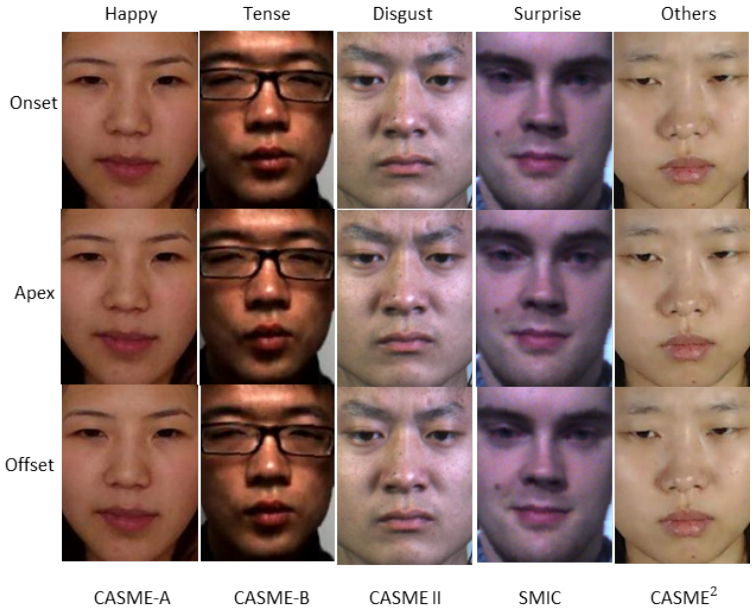


FIGURE 5.5: Example samples from CASME-A,CASME-B, CASME II, SMIC and (CAS(ME)^2) database. Three frames of micro-expressions are presented, including onset, apex and offset.

Algorithm 1: A Micro-Expression Apex Detection Procedure

Input : A sequence containing M frames.

Output : Apex of the micro-expression.

Procedure

1: Track 49 facial key-points for each frame.

2: Register faces for keeping eyes in horizontal line. Since locations of facial key-points have already changed after registration, a re-tracking operation is needed for cropping faces.

3: Crop and mask faces.

4: Divide the face into two parts : the forehead and the mouth part.

5: Compute the optical flow for each part.

6: Compute the sum of the magnitude of optical flow for each part in each frame and obtain two sequences. Sum these two sequences.

7: Apply the Gaussian smoothing operation ($\sigma = 2$) on the sequence in step 6.

8: Detect the micro-expression apex corresponding to the maximum of the filtered sequence.

9: If no apex is found, set the middle frame of the sequence as the apex.

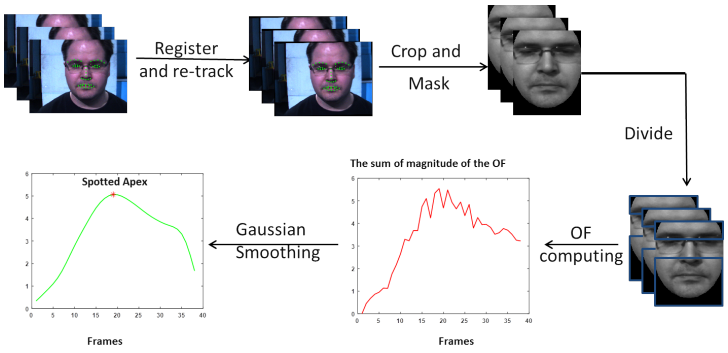


FIGURE 5.6: Illustration of the algorithm 1.

5.4.2 Baseline features

The proposed FMBH feature is compared with four baseline features in micro-expression recognition, i.e., the OF, MB, HOF and MBH :

- OF : The optical flow developed by Black et al. [BA96] is used. The apex and onset image of each sequence are used to compute the optical flow. The computed optical flows are resized to $N \times N$. The default parameter N for our experiments is 32. Thus, the feature dimension of the OF is $2 \times N \times N$.
- MB : The MB descriptor separates optical flow $\mathcal{F} = (\mathfrak{p}, \mathfrak{q})$ into its horizontal component $\nabla\mathfrak{p} = (\mathfrak{p}_x, \mathfrak{p}_y)$ and vertical component $\nabla\mathfrak{q} = (\mathfrak{q}_x, \mathfrak{q}_y)$, such that the feature dimension is twice of the dimension of OF up to $4 \times N \times N$.
- HOF : In order to compute the HOF descriptor [DTS06], the image domain is decomposed into $m \times n$ non-overlapping blocks, where m and n are fixed parameters. Each block can be indexed by $(\imath, \jmath)$, $1 \leq \imath \leq m$, $1 \leq \jmath \leq n$. For each block $(\imath, \jmath)$, we can generate a histogram of the optical flow features with respect to P bins, which is similar to the construction of the FMBH descriptor in Eqs. 5.9 and 5.10. By concatenating histograms extracted block by block, we thus obtain the HOF descriptor with length $P \times m \times n$.
- MBH : As depicted in Section 5.3.1, the MBH descriptor [DTS06] is established by the gradient vector fields of $\mathfrak{p}$ and $\mathfrak{q}$. We first decompose the image domain into $m \times n$ blocks. For each gradient vector field, a combinational histogram is obtained by concatenating the histograms respectively from each gradient vector fields of $\mathfrak{p}$ and $\mathfrak{q}$. Then the final MBH descriptor can be computed by concatenating these combinational histograms using the same procedure as the HOF descriptor, where

the length of the final historgram is $2 \times P \times m \times n$.

In order to prevent over-fitting, the principal component analysis is performed on all the features and the SVM is applied for training and testing. For the HOF, MBH and FMBH, the number of principal components of feature dimensions are kept by the PCA from 85% to 99% with a step length of 2% for comparing their influence. As the feature length of the OF and MB are much higher than the HOF, MBH and FMBH, their feature dimensions are kept from 50% to 95% with a step length of 5% for classification. We finally determine the remaining length of the feature dimensions in terms of the recognition rate after PCA operation. In other words, the length of the feature dimensions will be chosen if the corresponding recognition rate is the highest one among all the candidates of feature dimension lengths.

Before the optical flow computation, face images are masked for removing inhomogeneous background and hairs region which may influence the recognition results. Different masks are applied depending on the image resolution in the three datasets. Procedure of building faces masks are provided in Algorithm 2 and Fig. 5.7, where the left image shows the average face in CAS(ME)^2 dataset, the middle image illustrates ellipse curve on average face and the right image shows obtained binary mask from the inner region of the ellipse.

Algorithm 2: Face Mask Generation

Input : All cropped faces in a dataset.

Output : A face mask.

Procedure

1: Compute an average face by calculating the average of all cropped faces in database.
2: Exploit the function "imellipse" (the function "imellipse" is used for creating draggable ellipse in MATLAB) on the average face and drag ellipse to a proper size manually. Thus, a binary mask is obtained.
3: Apply the binary mask on faces.

Some examples are shown in Fig. 5.8 to visualize this image pre-processing procedure.

For classification, a non-linear SVM [CL11] with an RBF kernel (see Eq. 20) is employed. Two important parameters are set in the SVM, which are γ and the penalty parameter C. For the RBF kernel, multiples (e.g. 2^k for $k \in [-8, 8]$) of the default value are used as search range in a grid-search using cross-validation to determine C and γ. The

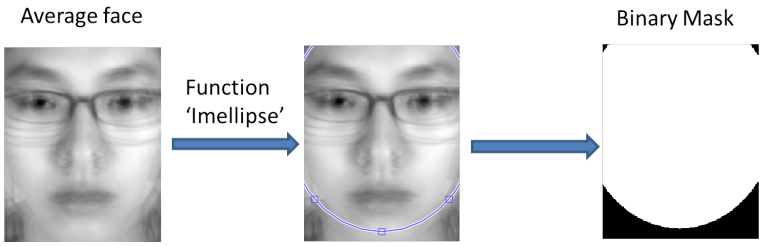


FIGURE 5.7: The process of creating a binary mask in CAS(ME)^2 database.

highest recognition rate is selected which corresponds to the optimal set of parameters.

The recognition experiments are conducted by applying the leave-one-subject-out (LOSO) where one subject is used as the test set and the others are used as the training set in each loop. The same parameters are used on the three datasets. We set $N = 32$, the values of m and n range from 1 to 10.

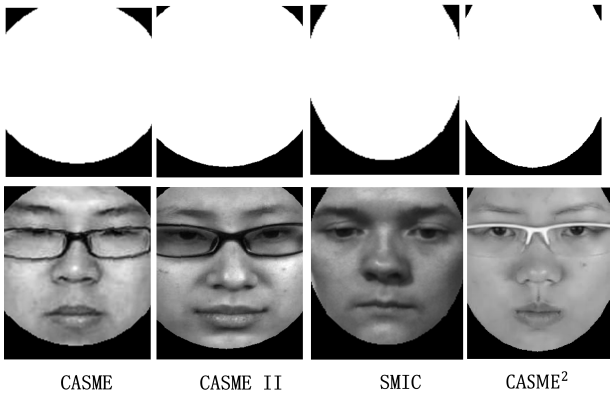


FIGURE 5.8: Illustration of masked samples from CASME, CASME II and SMIC database, respectively. The first line represents three masks and the second line provides samples masked by the corresponding mask.

5.5 Experimental results

In this section, the FMBH is evaluated comparatively with other motion descriptors on the three datasets. We first discuss the performance of different descriptors for different number of bins in Section 5.5.1. Section 5.5.2 presents the results of the FMBH using different blocking. In Section 5.5.3, the FMBH is first compared with four baselines, i.e., the OF, MB, HOF, MBH, and then compared to state-of-the-art results.

5.5.1 Comparison of performance of descriptors under different number of bins

The recognition rates obtained from motion descriptors (i.e., HOF, MBH and FMBH) under different number of bins are compared in Tables 5.1, 5.2, 5.3 and 5.4, respectively. We report the performance of each individual descriptor with respect to different number of bins. Through all the bin modes, the average recognition rates for the FMBH descriptor achieves 58.33% on CASME, 64.29% on CASME II, 68.34% on SMIC and 72.05% on CAS(ME)^2, which outperforms the performance of the HOF and MBH descriptors. It should not be surprised that the results in CASME II and SMIC datasets are better than those in CASME, because CASME II, SMIC and CAS(ME)^{2}datasets were filmed without illumination variance. Specifically, we take the results derived from 2 bins, i.e., $P = 2$, as an example. We observe that the FMBH descriptor achieves 6.66%, 1.63%, 1.83% and 0.54% higher rates than the MBH descriptor on the datasets of CASME, CASME II, SMIC and CAS(ME)^2, respectively. We also observe that both the FMBH and MBH descriptors are more discriminative and stable for micro-expression recognition than the optical flow.

From Tables 5.1, 5.2, 5.3 and 5.4, it can be observed that the number of bins and the respective rotations give rise to a slight influence in performance. Different descriptors achieve their best performance at different values of P. The proposed FMBH descriptor achieves the best performance using $P = 3$ under Mode 4 at CASME, $P = 4$ at SMIC, $P = 8$ at CASME II and $P = 7$ at CAS(ME)^2. However, for the dataset CASME II, our descriptor achieves the third highest recognition rate using $P = 4$. Such an observation can give a useful suggestion when choosing the number of bins for practical applications.

Table 5.1: Comparison of recognition rates of descriptors under different number of bins in CASME dataset

Methods	Number of bins P												Mean
	$P = 2$	$P = 3$				$P = 4$		$P = 5$	$P = 6$	$P = 7$	$P = 8$	$P = 9$	
		Mode 1	*Mode 2*	*Mode 3*	*Mode 4*	*Mode 1*	*Mode 2*						
HOF	60%	56.67%	58%	60%	55.33%	58%	58%	58.67%	58.67%	55.33%	54%	54%	57.23%
MBH	54.67%	56%	58.67%	59.33%	60.67%	58.67%	56.67%	58.67%	56.67%	56%	56%	57.33%	57.44%
FMBH	**61.33%**	59.33%	59.33%	59.33%	**61.33%**	56.67%	58.67%	58.67%	56%	56.67%	56.67%	56%	58.33%

Table 5.2: Comparison of recognition rates of descriptors under different number of bins in CASME II dataset

Methods	Number of bins P												Mean
	$P=2$	$P=3$				$P=4$		$P=5$	$P=6$	$P=7$	$P=8$	$P=9$	
		Mode 1	*Mode 2*	*Mode 3*	*Mode 4*	*Mode 1*	*Mode 2*						
HOF	64.63%	60.98%	61.79%	65.45%	61.38%	63.01%	64.23%	59.35%	61.38%	61.79%	63.01%	63.01%	62.50%
MBH	67.07%	66.67%	63.82%	66.26%	62.60%	64.23%	65.04%	64.63%	64.23%	64.23%	62.60%	62.60%	63.58%
FMBH	68.7%	63.82%	64.23%	62.20%	64.63%	66.67%	62.60%	63.01%	62.20%	63.41%	**69.11%**	60.98%	64.29%

Table 5.3: Comparison of recognition rates of descriptors under different number of bins in SMIC dataset

Methods	Number of bins P												Mean
	$P=2$	$P=3$				$P=4$		$P=5$	$P=6$	$P=7$	$P=8$	$P=9$	
		Mode 1	*Mode 2*	*Mode 3*	*Mode 4*	*Mode 1*	*Mode 2*						
HOF	61.59%	67.07%	67.68%	**71.95%**	64.63%	64.63%	70.12%	67.68%	66.64%	62.2%	63.41%	61.59%	65.76%
MBH	67.68%	65.24%	67.68%	69.51%	69.51%	68.29%	69.51%	65.85%	71.34%	66.46%	68.29%	67.68%	68.08%
FMBH	69.51%	65.85%	71.34%	67.07%	68.29%	**71.95%**	70.12%	66.46%	65.24%	68.90%	70.12%	65.24%	68.34%

Table 5.4: Comparison of recognition rates of descriptors under different number of bins in CAS(ME)^{2}dataset

Methods	Number of bins P												Mean
	$P=2$	$P=3$				$P=4$		$P=5$	$P=6$	$P=7$	$P=8$	$P=9$	
		Mode 1	*Mode 2*	*Mode 3*	*Mode 4*	*Mode 1*	*Mode 2*						
HOF	62.46%	64.71%	64.99%	60.78%	58.26%	63.03%	65.55%	61.34%	61.34%	64.99%	64.43%	62.46%	62.86%
MBH	71.43%	71.15%	71.99%	70.87%	72.55%	70.31%	73.11%	71.99%	72.83%	70.03%	70.59%	71.99%	71.52%
FMBH	72.55%	72.55%	72.27%	71.43%	72.27%	72.27%	71.71%	73.39%	71.15%	**73.67%**	70.31%	71.15%	72.06%

5.5.2 Comparison of different blocking

We investigate the impact of face blocking on the recognition rate. Table 5.5 illustrates the performance of the FMBH on SMIC with a bin number of 4 under Mode 1. The grid levels range from 1 to 10. The best performance is observed as 71.95% with 9×4 blocks. The performance of the FMBH on the whole face (1×1 block size) is 38.41%, much lower than 71.95%, which demonstrates that observing information spatially has an influence on recognition rate.

TABLE 5.5: Recognition rates with respect to different block sizes. The m and n represent the horizontal and vertical grid level.

Grid level		m									
		1	2	3	4	5	6	7	8	9	10
n	1	38,41%	37,20%	51,22%	46,95%	47,56%	53,66%	49,39%	46,95%	54,27%	51,83%
	2	49,39%	45,73%	46,95%	50,00%	52,44%	53,66%	53,05%	57,32%	51,22%	51,83%
	3	53,66%	52,44%	53,66%	55,49%	53,66%	56,10%	54,88%	57,93%	53,66%	57,32%
	4	56,10%	59,15%	62,80%	64,02%	60,98%	62,80%	62,80%	60,37%	60,37%	61,59%
	5	65,85%	62,20%	63,41%	67,07%	62,80%	60,98%	62,20%	62,20%	60,98%	61,59%
	6	62,80%	62,20%	63,41%	64,63%	62,20%	62,80%	62,80%	61,59%	62,20%	63,41%
	7	62,80%	61,59%	63,41%	67,07%	60,98%	62,20%	62,20%	62,80%	60,98%	61,59%
	8	60,37%	62,80%	67,07%	64,02%	62,20%	63,41%	62,80%	64,02%	62,80%	63,41%
	9	67,07%	66,46%	65,24%	**71,95**%	65,24%	62,80%	62,80%	63,41%	59,76%	62,20%
	10	62,80%	62,80%	63,41%	69,51%	64,63%	61,59%	64,02%	62,20%	60,37%	60,37%

5.5.3 Comparison of baseline descriptors with state-of-the-arts

In this section, we first present the comparison results of the FMBH to the existing motion descriptors, i.e., the OF, MB, HOF and MBH, and then compare our results to the state-of-the-art methods (i.e., LBP-TOP [ZP07], STLBP-IIP [HWL$^+$16], DiSTLBP-IIP [HWL$^+$16], HIGO/HOG+XOT/TOP [LHM$^+$15], STCLQP [HZH$^+$16a], MDMO [LZY$^+$16], Bi-WOOF [LSPW16] and FDM [XZW17]) in recent literature. Features in these state-of-the-art methods are classified into two categories : texture and motion features, where texture features include LBP-TOP, STLBP-IIP, DiSTLBP-IIP, HIGO/HOG+XOT/TOP and STCLQP, while motion features contain the MDMO, Bi-WOOF and FDM. The results are displayed in Table 5.6. For the purpose of fair comparisons, the recognition rates associated to the HOF, MBH and FMBH descriptors are obtained by choosing the respective highest rate through all the four datasets, which can also be seen from Tables 5.1 to 5.3.

TABLE 5.6: Recognition rates comparison with the state-of-art methods on three datasets by using leave-one-subject-out protocol. The bold means the highest recognition rate and * means that we directly extract the results from the reference paper.

Methods	CASME	CASME II	SMIC	CAS(ME)^2
OF	55.33%	64.63%	66.46%	68.91%
MB	60%	63.41%	69.51%	73.11%
HOF	60%	65.45%	**71.95%**	65.55%
MBH	60.67%	67.07%	71.34%	73.11%
LBP-TOP* [QWY$^+$17]	/	/	/	40.83%
STLBP-IIP* [HWL$^+$16]	59.06%	62.75%	60.37%	/
DiSTLBP-IIP* [HWL$^+$16]	64.33%	64.78%	63.41%	/
HIGO/HOG+XOT/TOP* [LHM$^+$15]	/	57.49%	65.24%	/
STCLQP* [HZH$^+$16a]	57.31%	58.39%	64.02%	/
MDMO* [LZY$^+$16]	**68.86**%	67.37%	/	/
Bi-WOOF* [LSPW16]	/	58.85%	62.20%	/
FDM* [XZW17]	56.14%	45.93%	54.88%	/
Proposed FMBH	61.33%	**69.11**%	**71.95**%	**73.67**%

Among all the baseline motion descriptors (OF, MB, HOF, MBH, FMBH), the FMBH feature achieves the best recognition rate (61.33% on CASME, 69.11% on CASME II, 71.95% on SMIC and 73.67% on CAS(ME)^2), which shows that building the relations between $\theta_{\mathfrak{p}}$ and $\theta_{\mathfrak{q}}$ as well as the relations between $M_{\mathfrak{p}}$ and $M_{\mathfrak{q}}$ are useful for improving micro-expression recognition rates. FMBH is then followed by the MBH and HOF.

One can observe from Table 5.6 that the FMBH outperforms the state-of-the-art methods evaluated upon these datasets except CASME in which the MDMO yields the highest recognition rate. On CASME II, the FMBH outperforms the DiSTLBP-IIP by over 4%. On SMIC, it is over 6% higher than the method in [LHM$^+$15]. Over CAS(ME)^2, the recognition rate of our method outperforms the LBP-TOP operator. Specifically, the rate of our method achieves 32.84% higher than LBP-TOP. Meanwhile, the MB, HOF, MBH achieve better performance than that of methods in [LHM$^+$15, HZH$^+$16a] on CASME II, CAEME^2 and SMIC. This is because the motion descriptors have a stronger ability of capturing small micro-movements than those texture descriptors. The facial dynamics map (FDM) [XZW17] based on optical flow estimation works poorly in micro-expression recognition. This demonstrates that our method using two frames (onset & apex) to extract motion information rather than the whole sequence can improve the accuracy of micro-expression recognition.

We further compare the confusion matrices of OF, MB, HOF, MBH and FMBH when they obtain the best recognition rates on CASME II in Table 5.7. The results show that the FMBH achieves higher accuracy on Disgust. Unfortunately, all motion features make much false classification of Repression to Others and also of Disgust to Others. This can be explained that Others class contains some confused micro-expressions similar to Disgust or Repression. The confusion matrices of FMBH on CASME, SMIC and CAS(ME)^{2}are displayed in Table 5.8. From these comparisons, we see that FMBH has a promising ability to recognize micro-expressions on the three databases.

Table 5.7: The confusion matrices of the (a) OF, (b) MB, (c) HOF, (d) MBH, (e) FMBH on the CASME II database at the best recognition rate, by the LOSO cross-validation

(a) OF

		Ground truth				
		Disgust	Happy	Repression	Surprise	Others
Prediction	Disgust	61.9%	9.38%	3.7%	0%	20.2%
	Happy	1.59%	43.75%	14.81%	12%	5.05%
	Repression	1.59%	12.5%	51.85%	0%	0%
	Surprise	0%	3.13%	0%	76%	1.01%
	Others	34.92%	31.25%	29.63%	12%	73.74%

(b) MB

		Ground truth				
		Disgust	Happy	Repression	Surprise	Others
Prediction	Digust	65.08%	9.38%	18.52%	12%	18.18%
	Happy	3.17%	56.25%	18.52%	0%	8.08%
	Repression	1.59%	3.13%	25.93%	0%	2.02%
	Surprise	0%	3.13%	0%	80%	1.01%
	Others	30.16%	28.13%	37.04%	8%	70.71%

(c) HOF

		Ground truth				
		Disgust	Happy	Repression	Surprise	Others
Prediction	Digust	52.38%	12.5%	7.41%	0%	18.18%
	Happy	7.94%	62.5%	11.11%	4%	7.07%
	Repression	0%	6.25%	62.96%	8%	2.02%
	Surprise	0%	6.25%	3.7%	76%	0%
	Others	39.68%	12.5%	14.81%	12%	73.73%

(d) MBH

		Ground truth				
		Disgust	Happy	Repression	Surprise	Others
Prediction	Disgust	66.67%	6.25%	0%	4%	11.11%
	Happy	0%	65.63%	22.22%	8%	8.08%
	Repression	0%	0%	25.93%	0%	6.06%
	Surprise	0%	3.13%	0%	88%	1.01%
	Others	33.33%	25%	51.85%	0%	73.74%

(e) FMBH

		Ground truth				
		Disgust	Happy	Repression	Surprise	Others
Prediction	Disgust	73.02%	15.63%	3.7%	4%	17.17%
	Happy	4.76%	62.50%	7.41%	8%	7.07%
	Repression	0%	6.25%	51.85%	4%	4.04%
	Surprise	0%	3.13%	3.7%	80%	1.01%
	Others	22.22%	12.5%	33.33%	4%	70.71%

The computation time for the optical flow estimation on two frames with size 250×200 is 12.02×10^3 ms. Based on the computed optical flow, the computation time for the MB operator, the HOF operator, the MBH operator and the FMBH operator are 3.04 ms, 3.47 ms, 5.32 ms and 4.3 ms, respectively.

TABLE 5.8: The confusion matrices of the FMBH on the (a) CASME, (b) SMIC, (c) CAMSE^2 database at the best recognition rate, by the LOSO cross-validation

(a) CASME

		Ground truth			
		Disgust	Repression	Surprise	Tense
Prediction	Disgust	35%	0%	0%	7.94%
	Repression	7.5%	44.83%	5.56%	7.94%
	Surprise	2.5%	3.45%	83.33%	4.76%
	Tense	55%	51.72%	11.11%	79.37%

(b) SMIC

		Ground truth		
		Positive	Negative	Surprise
Prediction	Positive	80.39%	21.43%	6.98%
	Negative	11.76%	64.29%	18.6%
	Surprise	7.84%	14.29%	74.42%

(c) CAS(ME)^2

		Ground truth			
		Positive	Negative	Surprise	Others
Prediction	Positive	77.42%	9.52%	12%	28.05%
	Negative	4.03%	77.78%	4%	7.32%
	Surprise	8.1%	23.8%	64%	1.22%
	Others	17.74%	10.32%	20%	63.41%

5.6 Conclusion

This chapter introduces and comparatively evaluates a new feature extraction method for the application of micro-expression recognition. The main contribution lies at the construction of facial features established based on differential optical flow vector fields. For this purpose, a nonlinear mapping is defined to establish a fusion rule that combines the respective gradient vector fields of the two optical flow components. The histograms of the proposed features are extracted from the Frobenius norms of a Jacobian matrix derived from the optical flow. To evaluate and optimize the performance of the proposed features, we have also investigated the influence of different ways of bin construction depending on the number of bins and the initial orientation angle of the first bin. The experiments conducted on four well-known micro-expression datasets show that the proposed method achieves promising results. Next chapter will summarize this thesis and present the future work with respect to micro-expressions analysis.

Chapter 6

Conclusion and Perspectives

This thesis is devoted to the micro-expression detection and recognition in videos. After extensive analysis of the state-of-the-art methods of the target topic, we have proposed some feature extraction methods for micro-expressions description. Each proposal is evaluated and tested upon well-known databases.

The main contributions are summarized hereafter and then we discuss the possible future work for their extension.

- **Framework of micro-expressions detection based on the IP feature**

  The detection method based on the IP feature is proposed to reduce the computation complexity for feature extraction. In addition, an automatic reference frame selection algorithm is developed for the purpose of reducing the accumulating errors along the sequence. For traditional texture features such as the LBP and the HOG features, and the motion feature like the optical flow, the computation time required is very expensive, limiting the possible applications in real-time detection. Experiments show that our proposed method can obtain better or comparable results but requiring much less computation time than those traditional texture and motion features.

- **Framework of micro-expressions detection based on the geometrical feature**

  The second method proposed for micro-expressions detection is based on the geometrical feature of key-points detected in the face. This method aims at comparing differences between frames along a sequence by extracting euclidean distance among facial key points. Traditional texture or motion features are extracted from

the faces cropping, while the proposed feature extraction method involves only facial key points such that the cropping faces operation is no longer required. Thus, errors given arise by the faces cropping step and the time consumed on the cropping faces can be avoided. By comparing with state-of-the-art methods, the experiments conducted on the SMIC dataset show that the proposed method achieves best results.

- **Framework of micro-expressions recognition based on the motion feature**

  For micro-expressions recognition, a fusion motion boundary histograms (FMBH) is proposed by combing both the horizontal and the vertical components of the differential of the optical flow. A relation is established between two gradients of the differential of the optical flow through a nonlinear mapping. Like the motion boundary histograms (MBH), the FMBH can reduce the unexpected motions caused by residual mis-registration. Nevertheless, the relative motion can be captured. Moreover, different number of orientation bins and its rotation modes are investigated in our experiments. The experiments conducted on four well-known micro-expression datasets show that the proposed method achieves promising results.

Future Work

Some future works can be derived from this thesis. First, it is valuable to further investigate the geometrical feature for micro-expressions detection and recognition. To be honest, the geometrical feature proposed in this thesis is not complete and must be improved. For example, the width between eyebrows are not included in our method, as well as the angle of the eyebrows, eyes and mouth corner. The geometrical distances between facial key points are Euclidean distances without any direction information. Thus, adding the direction information may improve the detection and recognition performance. Since blinking of eyes may cause rapid movements in the skin around eyes and eye brows, traditional detection methods based on features such as the LBP, HOG, OF and IP have been proved to fail distinguishing the movements of the eyes with emotions from the natural eyes blinks [MZP14b, DYC+14, LKR17, WWQ+17]. While the geometrical distance is not influenced by movements in the skin such that the detection method based on the geometrical feature is capable of removing the influence of natural eye blinks in videos. While the traditional detection methods based on features such as the LBP, HOG, OF

and IP have been proved to fail distinguishing the movements of the eyes with emotions from the natural eyes blinks.

Second, it is worthwhile to develop more powerful faces alignment approaches. Since the movement of micro-expressions is subtle, any small non-mapping images caused by faces alignment or cropping operation may lead to wrong orientation and magnitude in the computation of motion features.

Third, it is meaningful to further study other motion features, or their combination. There exists other motion features such as the divergence and the curl, besides those motion features mentioned in Chapter 5.

Fourth, it is interesting to investigate the influence of different machine learning classifiers for micro-expressions detection and recognition. Traditional recognition methods focus on the extraction of features and gives little importance to the selection of machine learning classifiers. Recently, the deep learning is popularly used in computer vision. The recognition method using deep learning can be an alternative way for improving its performance.

Finally, more micro-expressions database should be built in the future. Most databases use young students or teachers who have never criminal experience, restricting the databases to analyze deception in real life, high-stake situation, or medical treatment. It is not possible to find a database for illumination related studies which require various illuminations. Moreover, approaches associated to occlusion are important because in real world, partial occlusion appears frequently. The system must be capable of recognizing micro-expressions despite occlusions by sunglasses, facial hairs, hands, scarves, etc. In general, it is difficult to create a database that will satisfy everyone's need. Publicly and freely available databases containing more samples, data under real deception environment, varying conditions of occlusion, varying lighting conditions, etc. are welcome for this work.

Appendix A : Support Vector Machine (SVM)

This appendix introduce a detailed description of the SVM, which is a discriminative classifier by a separating hyperplane. Suppose given a set of labeled training data, the SVM will output an optimal hyperplane which classifies testing data. This appendix firstly reviews both linear and non linear separable case, including how to find an optimal separating. Then, three widely used kernels of linear, polynomial and RBF are introduced in Section 6. Section 6 introduce problems of SVM in overfitting and error tolerance. SVM algorithm is detailed described in Section 6. The conclusion is presented at last.

The Linear Separable Case

A simple model of 2D linear classifier is shown as Figure 1.

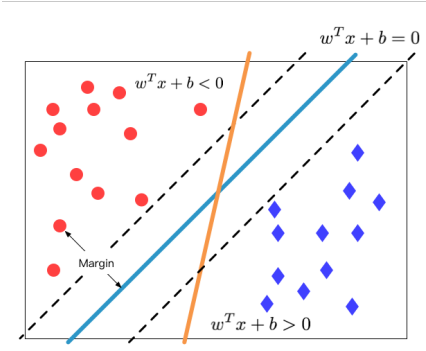


FIGURE 1: Linear Classifier Schematic diagram

A training dataset with n points is plotted on the image, which has a form of :

$$(\vec{x_1}, y_1), (\vec{x_2}, y_2), ..., (\vec{x_n}, y_n) \tag{1}$$

where $\vec{x_i}$ is the p-dimensional vector and y_i is 1 or -1 indicating the class for each point. In the image, we can see orange and blue lines can both identify two groups of data, but how can we know which linear classifier is better ? To solve this problem, maximum margin classifier was proposed. The margin is defined as the geometrical distance from each point to hyperplanes, which is $\gamma_i = \frac{w^T x_i + b}{\|w\|}$. For linear separable problem, all the parallel linear classifier with no data point between each other, could be classified into same group. And the borders of such a group are defined as $w^T x_i + b = k$ and $w^T x_i + b = -k$. Through choosing the value of $\|w\|$, k is also ranging from 0 to ∞, then we set $k = 1$ by scaling the $\|w\|$. The maximum margin classifier for each group just lies on the middle of the two borders, and the distance between the two hyperplane is $\frac{2}{\|w\|}$. As no data point falls in the margin, the limitation to two borders is :

$$w^T x_i + b \geq 1 \quad \forall y_i = 1$$

and

$$w^T x_i + b \leq -1 \quad \forall y_i = -1$$

Then the final optimization problem becomes

$$\max \frac{2}{\|w\|}, \ s.t. \ \ y_i(w^T x_i + b) \geq 1, for \ i = 1, \ldots, n$$

Additionally, the points most close to the maximum margin classifier must lie on the both borders. These points are called support vectors.

As

$$\max \frac{2}{\|w\|} \iff \min \frac{1}{2}\|w\|^2$$

the maximum margin problem could be converted to Convex optimization and solved by Quadratic Programming.

Another popular solution to this problem is Lagrange duality. The Lagrangian function is constructed as below :

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1] \tag{2}$$

Then set $\theta(w) = \max \mathcal{L}(w, b, \alpha) \ s.t. \ \ \alpha_i \geq 0$. If all limitations are satisfied, then $\theta(w) = \frac{1}{2}\|w\|^2$, otherwise $\theta(w) \to \infty$. Now, our optimization problem has been transformed to :

$$\min_{w,b} \theta(w) = \min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w, b, \alpha) \tag{3}$$

As this problem satisfies Karush-Kuhn-Tucker conditions (KKT) [DFKS11], we have :

$$\min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w, b, \alpha) = \max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w, b, a) \tag{4}$$

To get $\min_{w,b} \mathcal{L}(w, b, a)$, it should have :

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \tag{6}$$

which is equivalent to :

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{7}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{8}$$

Back to the original equation 2 :

$$\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \tag{9}$$

The final form of the optimization problem is :

$$\max_{\alpha} (\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i) \tag{10}$$

under the limitation of

$$\alpha_i \geq 0, i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

This is a dual variable problem and could be solved by Sequential Minimal Optimization, which is proposed by John C. Platt [Pla98].

The Non Linear Separable Case

In most of real situations, the data is not linearly separable in raw space. To solve this non-linear separable problem, a possible solution is to project these data onto a higher dimension.

Let ϕ be a map from low dimension to high dimension. According to Eq.(7), we have the hyperplane equation as below :

$$\begin{aligned} f(x) &= (\sum_{i=1}^{n} \alpha_i y_i x_i)^T x + b \\ &= \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \end{aligned}$$

where $\alpha_i = 0$ for all the non support vector.

Then in new high dimension space, the hyperplane should be :

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b \tag{11}$$

The new optimization problem is converted to :

$$\max_{\alpha} (\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle) \tag{12}$$

under the limitation of

$$\begin{aligned} \alpha_i \geq 0, i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

Although ϕ transformation could help solve the non linear separable case, the computing load grows geometrically with the original data dimension, which makes it inefficient in real application. Thus the kernel trick was introduced to solve this problem.

The idea of "Kernel" comes from another mathematical tool in functional analysis, Reproducing kernel Hilbert space (RKHS), which is widely utilized even before the creation of SVM [RT01].

At first, an inner product should be defined as below :

Definition 1. Let $\mathcal{H}$ be a vector space over $\mathbb{R}$, if there is a map :

$$\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R} \tag{13}$$

satisfying the following three axioms for all vector $x, y, z \in \mathbb{R}$:

1. Conjugate symmetry : $\langle x, y \rangle_{\mathcal{H}} = \langle y, x \rangle_{\mathcal{H}}$
2. Linearity in the first argument : $\langle a_1 x_1 + a_2 x_2, g \rangle_{\mathcal{H}} = a_1 \langle x_1, g \rangle_{\mathcal{H}} + a_2 \langle x_2, g \rangle_{\mathcal{H}}$

3. Positive-definiteness : $\langle x, x\rangle_{\mathcal{H}} \geq 0$ and $\langle x, x\rangle_{\mathcal{H}} = 0 \iff x = 0$

Then it is said to be a inner product. A norm of any vector could be defined as its inner product with itself : $\|x\|_{\mathcal{H}} = \sqrt{\langle x, x\rangle_{\mathcal{H}}}$. A Hilbert space is a complete and separable space on which an inner product is defined.

According to Riesz representation theorem [Sch71], the definitions of Reproducing kernel Hilbert space and its kernel can be achieved [W$^+$99].

Definition 2. Let $\mathcal{H}$ be a Hilbert space of function $f : X \to \mathbb{R}$, and X is a non-empty set. For a fixed $x \in X$, bijective map $\delta_x : \mathcal{H} \to \mathbb{R}$, then $\delta_x : f \to f(x)$ is called the evaluation function at x. The Hilbert space $\mathcal{H}$ is called a RKHS when δ_x is continuous $\forall x \in X$

Definition 3. A function $K : X \times X \to \mathbb{R}$ is called a reproducing kernel if it satisfies :

(1) $\forall x \in X, K(\cdot, x) \in \mathcal{H}$

(2) $\forall x \in X, \forall f \in \mathcal{H}, \langle f, K(\cdot, x))\rangle_{\mathcal{H}} = f(x)$

The second property is called the reproducing property.

And in particular :

$$K(x, y) = \langle K(\cdot, x), K(\cdot, y)\rangle_{\mathcal{H}} \quad \forall x, y \in X \tag{14}$$

The kernel has several important properties :

(1) If it exists, reproducing kernel is unique.

(2) Reproducing kernels are positive definite.

(3) For every positive definite function K(x,y), there is a unique RKHS whose reproducing kernel is K Hilbert space, if and only if H has a reproducing kernel.

Here, we only show the demonstration of the kernel for any RKHS. Assume that $\delta_x \in \mathcal{H}'$ is a bounded linear functional. According to Riesz representation theorem (In a Hilbert space $\mathcal{H}$, all continuous linear functionals are of the form $\langle \cdot, g\rangle_{\mathcal{H}}$, for some $g \in \mathcal{H}$), there exists an element $f_{\delta_x} \in \mathcal{H}$ such that :

$$\delta_x f = \langle f, f_{\delta_x}\rangle_{\mathcal{H}}, \ \forall f \in \mathcal{H} \tag{15}$$

Define $K(x', x) = f_{\delta_x}(x')$, $\forall x, x' \in X$. Then, $K(\cdot, x) = f_{\delta_x} \in \mathcal{H}$, and $\langle f, K(\cdot, x)\rangle_{\mathcal{H}} = \delta_x f = f(x)$. Thus, K is the reproducing kernel.

Through the introducing of kernel trick, a map could be transformed to inner product in the original space. In our case, if the map from low dimension space to high dimension

space is $\phi(x_0) = K(\cdot, x)$, then

$$\langle \phi(x), \phi(y) \rangle = \langle K(\cdot, x), K(\cdot, y) \rangle_{\mathcal{H}} = K(x, y) \tag{16}$$

Thus the optimization problem return to the low dimension space under the form of :

$$\max_{\alpha} (\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x, y)) \tag{17}$$

under the limitation of

$$\alpha_i \geq 0, i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

But, for specific dimension transformation operator ϕ, K is always not easy to calculate and express. But as we do not need any information from high dimension space, usually we can confirm the kernel function firstly, then calculate the high dimension space. Additionally, there have been many useful kernels, which are already well studied and have good properties for the application. We will introduce these kernel function in next chapter.

SVM Kernels

There are several kernels being widely used in the kernel trick, such as : linear kernel, polynomial kernel, Gaussian kernel, RBF kernel etc.. Some of their mapping space has finite dimension, such as linear and polynomial kernels, while some are infinite, such as RBF kernels. Even more, researchers can build up their own kernel to fit their work better, on the basis of these simple kernels. In general, more target dimension will improve the classification accuracy for it provides more features, but will also led to over-fitting problems, whose introduction is in next section. As the result, how to choose the right kernel function and its corresponding parameters becomes a crucial part for the application of SVMs. There are two useful measurements to determine the kernel function [SS02]. The first is to test the classification accuracy with training dataset, and compare the results from different kernels. The other one is using the experiences from similar field. For example, second order Polynomial Kernels are demonstrated to be efficient for face recognition [OFG97, GLC00, Bur98].

Linear Kernels

Linear Kernels is the simplest kernel, which is given as below :

$$K(x, y) = x^T y + c \tag{18}$$

where x and y are vectors of features defined as above, and c is a parameter to trade off the influence of the order of the original space.

Linear kernel keeps the dimension of original space. In another way, it means the SVMs will collapse into Bayesian linear regression, and decrease its own time complexity from $O(N^3)$ to $O(N)$.

Polynomial Kernels

Polynomial kernel is a high-dimensional promotion of linear kernel. Although it has finite dimension, polynomial kernel has important position in that situation where overfit tend is obvious. For d-degree polynomials, the polynomial kernel is defined as below :

$$K(x, y) = (x^T y + c)^d \tag{19}$$

subject to

$$c \geq 0$$

where x, y, c have the same definitions as in linear kernel. Especially, this kernel is called homogeneous when $c = 0$.

RBF Kernels

(Gaussian) Radial basis function (RBF) kernel is defined as below :

$$K(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2}) \tag{20}$$

where x and y are vectors of features as above, $\|x - y\|$ is the Euclidean distance, and σ is a free parameter. The RBF kernel is the most used infinite kernel, for the distribution of real samples are usually Gaussian distribution in practice.

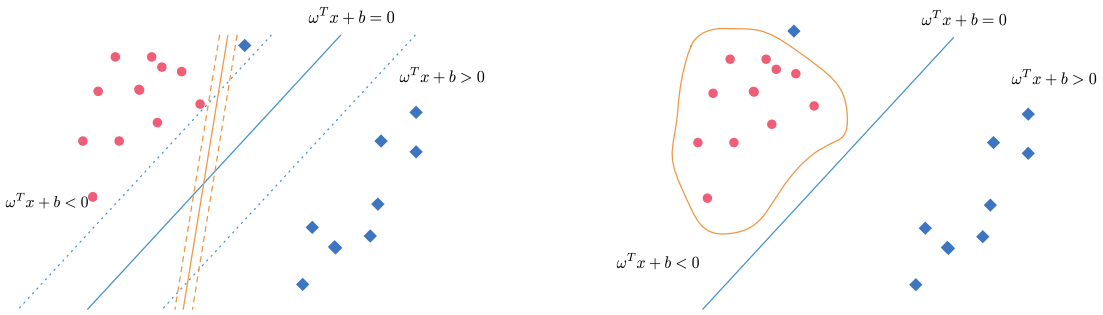


FIGURE 2: Linear Classifier with sample error

Overfitting and Error tolerance

In theory, any samples can be classified in high enough dimension space, but the classifier obtained may be inconsistent with the actual situation. What's worse is, if there are some errors or mistakes in training set, they will distort the classifier as in Fig. 2. In Fig. 2 left sub-figure, the blue line, who responses to the classifier, is shifted to the orange line. The new classifier has far less margin than the origin one. In Fig. 2 right sub-figure, the blue line is completely distorted into orange line, which means the classifier can only be solved from high dimension space, which increases the time complexity of the whole program. To avoid such shift and distortion which are usually called overfitting, the SVMs must have some kind of error tolerance, and its own accuracy as well. Then the soft margin SVM is developed to solve this problem.

From Eq. (6), the restriction is $y_i(w^T x_i + b) \geq 1, for\ i = 1, \ldots, n$. To tolerate the errors, relaxation variable $\xi_i \geq 0$ is introduced into the restrictions :

$$y_i(w^T x_i + b) \geq 1 - \xi_i, for\ i = 1, \ldots, n \tag{21}$$

which will permit some point to drop within the margin to achieve larger margin distance.

In the other hand, the error should be carefully controlled in a small amount to decrease accuracy loss. So a threshold function is added onto original objective function :

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{N} \xi_i \tag{22}$$

where parameter C is the level of user's tolerance to errors.

Subject to

$$\begin{aligned} y_i(w^T x_i + b) &\geq 1 - \xi_i, for\ i = 1, \ldots, n \\ \sum_{i=1}^{m} \alpha_i y_i &= 0 \end{aligned}$$

The Lagrangian function for this problem is :

$$\mathcal{L}(w, b, \alpha, \beta, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(w^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^{N}\beta_i\xi_i \tag{23}$$

When KKT condition is satisfied, the Lagrangian dual of above function is :

$$\max_{\alpha_i \geq 0, \beta_i \geq 0} \min_{w,b,\xi} \mathcal{L}(\omega, b, \alpha, \beta, \xi) \tag{24}$$

This problem is equivalent to :

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \tag{25}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \tag{26}$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = 0 \tag{27}$$

From Eq. (27), it is obvious that :

$$C = \alpha_i + \beta_i \tag{28}$$

Then all β_i could be replaced with $C - \alpha_i$ in Eq. (23) and Eq. (**??**). What's more, as $\alpha_i \geq 0, \beta_i \geq 0$, we have $0 \leq \alpha_i \leq C$.

$$\max_{0 \leq \alpha_i \leq C} \min_{w,b} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w^T x_i + b) - 1] \tag{29}$$

The method to solve this optimization problem is same as Eq.(4)-(24), the final form is :

$$\max_{\alpha} (\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i) \tag{30}$$

subject to

$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

the only difference is that the α_i has upper bound C.

And the kernel form of soft margin SVM is :

$$\max_{\alpha} (\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x, y)) \tag{31}$$

subject to

$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

The soft margin SVM is more suitable and practical for real classification problems. Through the C chosen, user could balance the error rate and margin distance. The only problem is that extra training samples and time are demanded for the parameter C.

Algorithm on solving SVMs

A normal SVM classifier induces a quadratic program (QP) as Eq. (31). As its quadratic form, the matrix has a number of elements quadratic to the sample number, which is hard to handle directly. Several attempts have been made since the propose of SVM, whose main idea is to break down the large problem into small parts.

The first algorithm, which was widely accepted by researchers, was proposed by Platt from Microsoft [Pla98]. This algorithm is named Sequential Minimal Optimization (SMO). After a short time, Shevade et al. improved the algorithm, and this version is directly named as Improved SMO (ISMO) [SKBM00, KSBM01].

SMO is an iterative algorithm to solve the upper problem, which sacrifices the time complexity for the space complexity. Its most different part from common QP iterative algorithm is that the SMO updates two variables for each step under the limitation of $\sum_{i=1}^{n} \alpha_i y_i = 0$.

For reducing the computational complexity, least squares support vector machine (LS-SVM) was invented by converting constraints from linear inequalities to linear equations, while this kind of conversion breaks the sparse characteristic of support vector. So it only simplified some specific issues [SV99].

Another progress improves the core mechanics and avoids to solve standard quadratic program. Lagrangian support vector machine (LSVM) is typical among this progress. LSVM requires a simpler iterative scheme and provides global resolution, while its main issue is coupling, which requests recalculation for each new introduced sample [MM01].

We introduce the algorithm and its implementation of SMO for soft margin SVM here.

Assume one step of the algorithm loop. The variable got updated is α_p and α_q, then other $\alpha_i (i \neq p, i \neq q)$ could be treated as constant, as the restriction $\sum_{i=1}^{n} \alpha_i y_i = 0$, we have :

$$\alpha_p y_p + \alpha_q y_q = -\sum_{i \neq p, i \neq q}^{n} \alpha_i y_i = Constant\, \eta \tag{32}$$

$$\alpha_p = \frac{\eta - \alpha_q y_q}{y_p} \tag{33}$$

From Eq. (31), set W as :

$$\begin{aligned} W &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x, y) \\ &= \alpha_p + \alpha_1 - \frac{1}{2} K(x_p, y_p) {\alpha_p}^2 - \frac{1}{2} K(x_q, y_q) {\alpha_q}^2 - y_p y_q K(x_p, x_q) \alpha_p \alpha_q - \\ &\quad y_p \alpha_p \sum_{i \neq p, i \neq q}^{n} \alpha_i y_i K(x_i, x_p) - y_q \alpha_q \sum_{i \neq p, i \neq q}^{n} \alpha_i y_i K(x_i, x_q) \end{aligned}$$

by taking Eq. (33) into above equation, we will get a univariate function $W(\alpha_q)$.

As our discussion above, the classifier with kernel can be expressed as :

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{34}$$

then

$$\sum_{i \neq p, i \neq q}^{n} \alpha_i y_i K(x_i, x_p) = f(x_p) - \alpha_p y_p K(x_p, x_p) - \alpha_q y_q K(x_q, x_p) - b \tag{35}$$

$$\sum_{i\neq p, i\neq q}^{n} \alpha_i y_i K(x_i, x_q) = f(x_q) - \alpha_p y_p K(x_p, x_q) - \alpha_q y_q K(x_q, x_q) - b \tag{36}$$

Set the error between calculation and real value as $E_i = f(x_i) - y_i$, and $\gamma = K(x_p, x_p) + K(x_q, x_q) - 2K(x_p, x_q)$,

Let :

$$\frac{\partial W}{\partial \alpha_q} = 0 \tag{37}$$

then

$$\alpha_q' = \alpha_q + \frac{y_q(E_p - E_q)}{\gamma} \tag{38}$$

For each step, as $0 \leq \alpha_i \leq C$, new variate α_q' must satisfy the restriction below :

$$L \leq \alpha_q' \leq H \tag{39}$$

where :

$$L = \begin{cases} \max(0, \alpha_q - \alpha_p) & y_q \neq y_p \\ \max(0, \alpha_q + \alpha_p - C) & y_q = y_p \end{cases} \tag{40}$$

$$H = \begin{cases} \min(C, C + \alpha_q - \alpha_p) & y_q \neq y_p \\ \min(C, \alpha_q + \alpha_p) & y_q = y_p \end{cases} \tag{41}$$

Then the final update function is :

$$\alpha_q' = \alpha_p + y_p y_q(\alpha_q - \alpha_q') \tag{42}$$

With the conclusion above, we can implement the SMO algorithm as below.

Algorithm 3: SMO for soft margin SVM

Input : C, kernel, kernel parameters, ξ

Result:

Initialize b and all α to 0;
while *KKT not satisfied* **do**
    repeat
        Find a sample e1 that violates KKT;
        Choose a second sample that violates KKT;
        $\alpha'_q = \alpha_q + \frac{y_q(E_p - E_q)}{\gamma}$;
        if $\alpha'_q > H$ **then**
            $\alpha'_q = H$;
        else if $\alpha'_q < L$ **then**
            $\alpha'_q = L$
        else
            $\alpha'_q = \alpha_p + y_p y_q(\alpha_q - \alpha'_q)$;
            Calculate E_i and b according to α'_p and α'_q;
        end
    until *The variation of* E_i *is less than the accuracy*;
end

Conclusion

This appendix introduce a supervised learning technique in the field of machine learning applicable to classification. The basic knowledge of the SVM is presented, along with three widely used kernels of linear, polynomial and RBF.

Appendix B : Comparison results of four dissimilarity metrics and Gaussian smoothing

In this section, the work associated to four dissimilarity metrics including the chi-squared distance, the kolmogorov-smirnow distance, the cosine distance and the Euclidean distance, and combined to the Gaussian smoothing is presented.

Dissimilarity analysis

There exists many other measures for the dissimilarity between two features. In our experiments, except the chi-squared distance, other four distance measures are also investigated in Section 16. Assuming $P = \{p_i\}$ and $Q = \{q_i\}$ are the two features at same length N :

- *Chi-squared distance* (CS), see Eq. (4.3).
- *Kolmogorov-Smirnow distance* (KS) [RTG00] :

$$\mathcal{D}_{KS} = \max_i(\|\hat{p_i} - \hat{q_i}), \quad i \in [1, N], \tag{43}$$

  where $\hat{p_i} = \sum_{j \leq i} p_i$ is the cumulative histogram of p_i, and similarly for q_i.
- *Cosine distance* (COS) [ZL03] :

$$\mathcal{D}_{COS} = 1 - \frac{\sum_{i=1}^{N} p_i q_i}{\sqrt{\sum_{i=0}^{N-1} p_i^2}\sqrt{\sum_{i=1}^{N} q_i^2}} \tag{44}$$

- *Euclidean distance* (EUC) [SB91] :

$$\mathcal{D}_{EUC} = \sum_{i=1}^{N} \sqrt{(p_i - q_i)^2}. \tag{45}$$

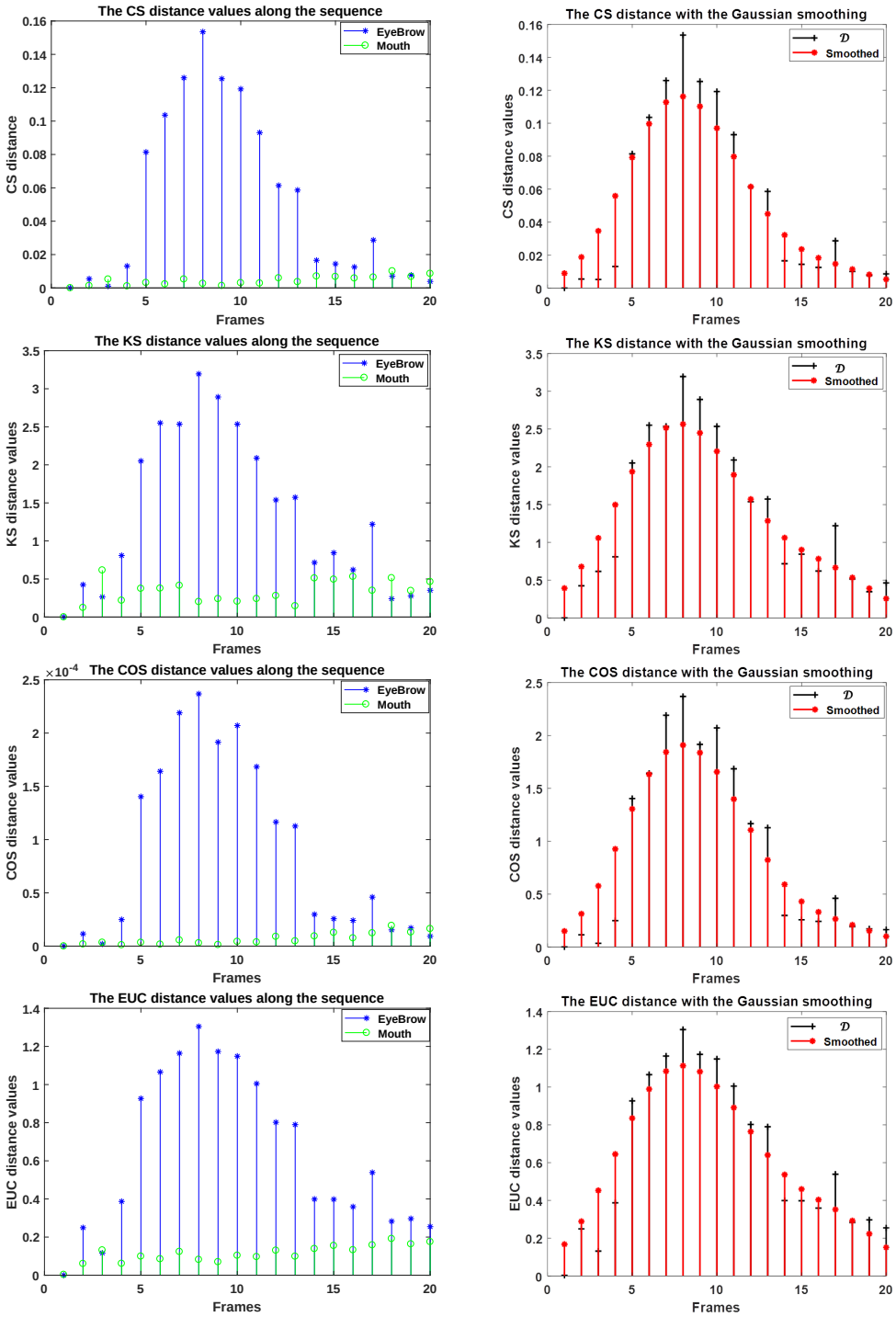


FIGURE 3: Comparison of different dissimilarity metrics. The left column represents the distance of the sequence which is labeled as "surprise"(file $s3\_sur\_02$ of the SMIC-HS database, by exploiting the CS, the KS, the COS and the EUC, respectively. The right column represents the curve with Gaussian Smoothing on $\mathcal{D}$ (refer to Section 4.4.2, Chapter 4), respectively. The σ of the Gaussian filter equals to 2.

Fig. 3 shows an example for the comparison of those four dissimilarity measure methods. The top row represents the distance of the sequence which is labeled as "surprise"(file $s3\_sur\_02$ of the SMIC-HS database), by exploiting the CS, the KS, the COS and the EUC, respectively. The second row represents the smoothed curves by applying the Gaussian Smoothing on the top row, respectively.

Experiments

For Gaussian smoothing, $\sigma = 2$, $\sigma = 3$ and $\sigma = 4$ are used for comparisons.

Fig. 4 to Fig. 7 show recognition rates under different combinations of methods applied to the SMIC-sub, HS, NIR and VIS dataset, respectively. (a) and (b) illustrate recognition rates using the RBF and polynomial kernels with four different dissimilarity metrics, respectively. (c) and (d) illustrate recognition rates using the RBF and polynomial kernels with the three standard deviations of the Gaussian filter. For (c) and (d), only the chi-squared distance is exploited because we find that in most cases, the recognition rates are higher by using the chi-squared distance than other four dissimilarity metrics as can be observed on (a) and (a).

Evaluation on SMIC-sub

As observed in Fig. 4 (a) and (b), the performance of chi-squared distance is better than other four dissimilarity metrics regardless using the RBF or polynomial kernel. Meanwhile, the performance of the cosine distance is worst among four dissimilarity metrics. For a more intuitive recognition rates comparisons between the RBF and polynomial kernels, as well as comparisons between the four dissimilarity metrics, we compute the mean recognition rate and the maximum recognition rate for each combination, see Table 1. The highest mean accuracy is achieved by combining the RBF kernel and chi-squared distance, yielding 82.32%. Both the combination of the chi-squared distance and the RBF kernel with a number of feature 19 (or $46 - 50$) and the combination of the chi-squared distance and the polynomial kernel with a number of feature 39 as well as $46 - 48$ reach the best result 85.53%.

From Fig. 4 (c) and (d), a slight improvement is observed by combining the RBF and the Gaussian filter with $\sigma = 3$ with a number of feature less than 16. Interpolating to a higher number of feature did not yield any improvement.

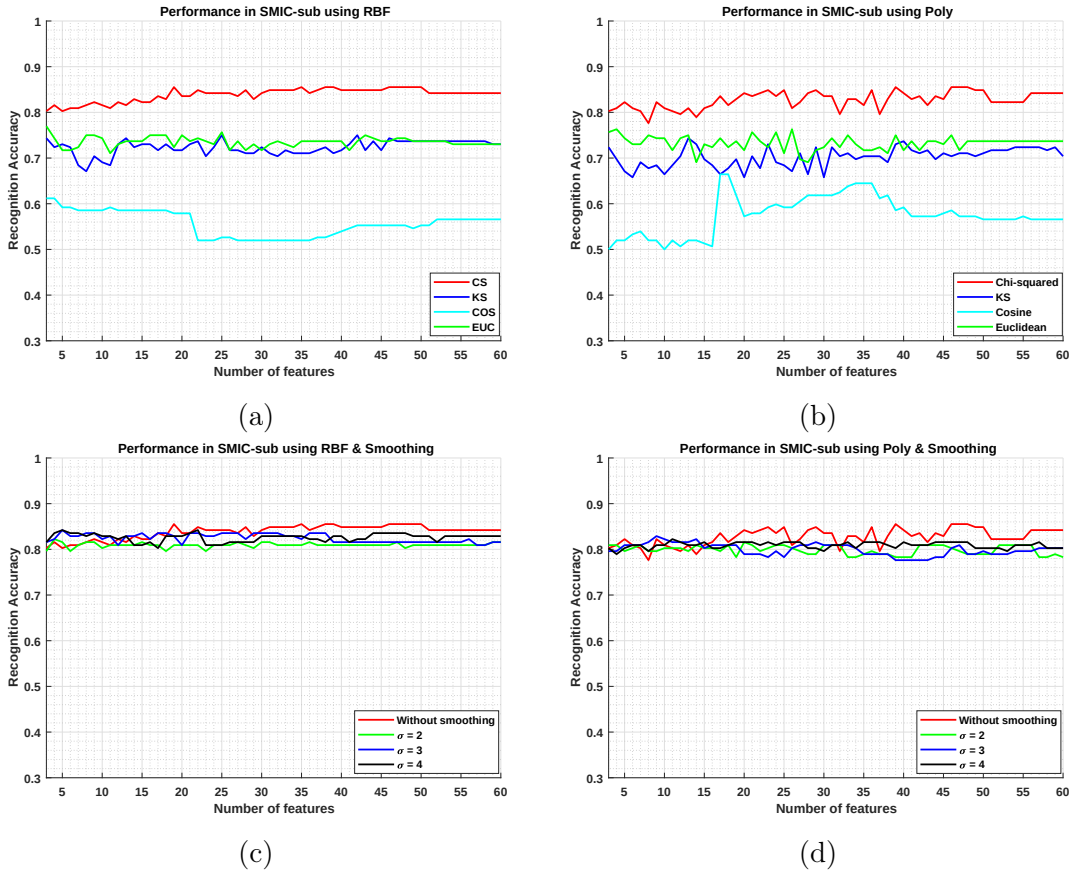


FIGURE 4: The results in the SMIC-sub dataset. **(a)** Comparison of four difference metrics regarding the influence of feature length on the recognition performance using the RBF kernel. **(b)** Comparison of four difference metrics regarding the influence of feature length on the recognition performance using the polynomial kernel. **(c)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the RBF kernel. **(d)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the polynomial kernel.

TABLE 1: Recognition rate comparisons between the RBF and polynomial Kernel as well as comparisons between four dissimilarity metrics in SMIC-sub.

Method	RBF				Poly			
	CS	KS	COS	EUC	CS	KS	COS	EUC
Mean	**83.32%**	72.32%	55.64%	73.55%	82.71%	67.70%	57.54%	73.28%
Max	**85.53%**	75.00%	61.18%	76.97%	**85.53%**	74.34%	66.45%	76.32%

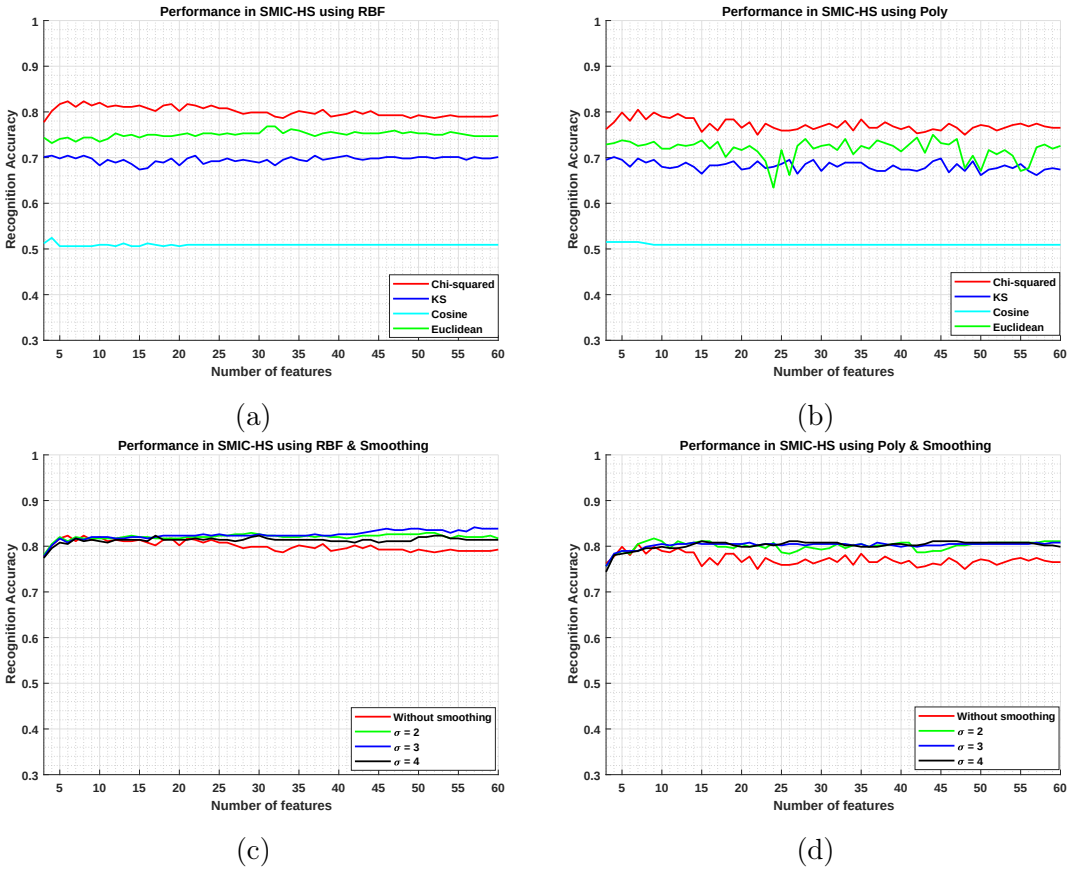


FIGURE 5: The results in the SMIC-HS dataset. **(a)** Comparison of four dissimilarity metrics regarding the influence of feature length on the recognition performance using the RBF kernel. **(b)** Comparison of four dissimilarity metrics regarding the influence of feature length on the recognition performance using the polynomial kernel. **(c)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the RBF kernel. **(d)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the polynomial kernel.

Evaluation on SMIC-HS

Fig. 5 (a) and (b) shows that the performance of chi-squared distance is still the best among four dissimilarity metrics regardless of using the RBF or polynomial kernel. Again, the performance of the cosine distance is worst among four dissimilarity metrics. Table. 2 shows a summary of the mean and the maximum recognition rate for each combination of method. Again, combining the RBF kernel and chi-squared distance yields the highest mean accuracy 80.09%. The combination of the chi-squared distance and the RBF kernel with a number of feature 6 (or 8) achieves the best result 82.32%.

Table 2: Recognition rate comparisons between the RBF and polynomial Kernel as well as comparisons between four dissimilarity metrics in HS dataset

Method	RBF				Poly			
	CS	KS	COS	EUC	CS	KS	COS	EUC
Mean	**80.89%**	69.59%	50.90%	75.05%	77.21%	68.13%	50.97%	71.81%
Max	**82.32%**	70.43%	52.44%	76.83%	80.49%	70.12%	51.52%	75.00%

From Fig. 5 (c) and (d), a significant improvement is observed by combining the RBF and the Gaussian filter with a number of feature larger than 24. For the HS dataset, Interpolating to a higher number of feature with the Gaussian filter did improve the performance. The highest recognition rate 84.15% is achieved by applying the Gaussian filter with $\sigma = 3$ and the RBF kernel with a number of feature 57.

Evaluation on SMIC-NIR

Fig. 6 (a) and (b) shows that the chi-squared distance slightly outperforms the euclidean distance combining the RBF and as well as the polynomial kernel. Table. 3 shows a summary of the mean and the maximum recognition rate for each combination of method. Again, combining the RBF kernel and chi-squared distance yields the highest mean accuracy 71.01%. The combination of the euclidean distance and the polynomial kernel with a number of feature 4 achieves the best result 75%.

From Fig. 6 (c) and (d), applying the Gaussian filter does not gain any improvements.

Evaluation on SMIC-VIS

Again, Fig. 7 (a) and (b) shows that the chi-squared distance outperforms other dissimilarity metrics combining the RBF and as well as the polynomial kernel. Table. 4

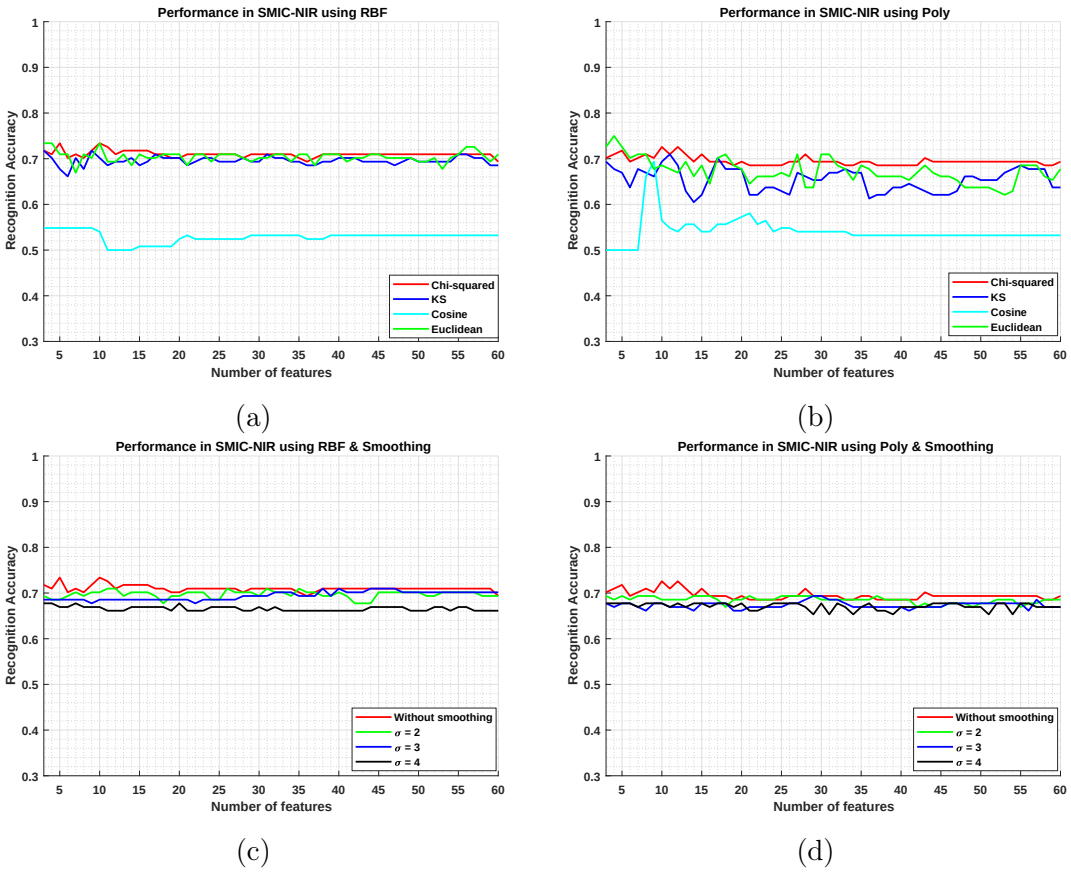


FIGURE 6: The results in the SMIC-NIR dataset. **(a)** Comparison of four difference metrics regarding the influence of feature length on the recognition performance using the RBF kernel. **(b)** Comparison of four difference metrics regarding the influence of feature length on the recognition performance using the polynomial kernel. **(c)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the RBF kernel. **(d)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the polynomial kernel.

TABLE 3: Recognition rate comparisons between the RBF and polynomial Kernel as well as comparisons between four dissimilarity metrics in NIR dataset

Method	RBF				Poly			
	CS	KS	COS	EUC	CS	KS	COS	EUC
Mean	**71.01%**	69.58%	52.85%	70.45%	69.51%	65.41%	54.21%	67.26%
Max	73.39%	71.77%	54.84%	73.39%	72.58%	70.97%	69.35%	**75.00%**

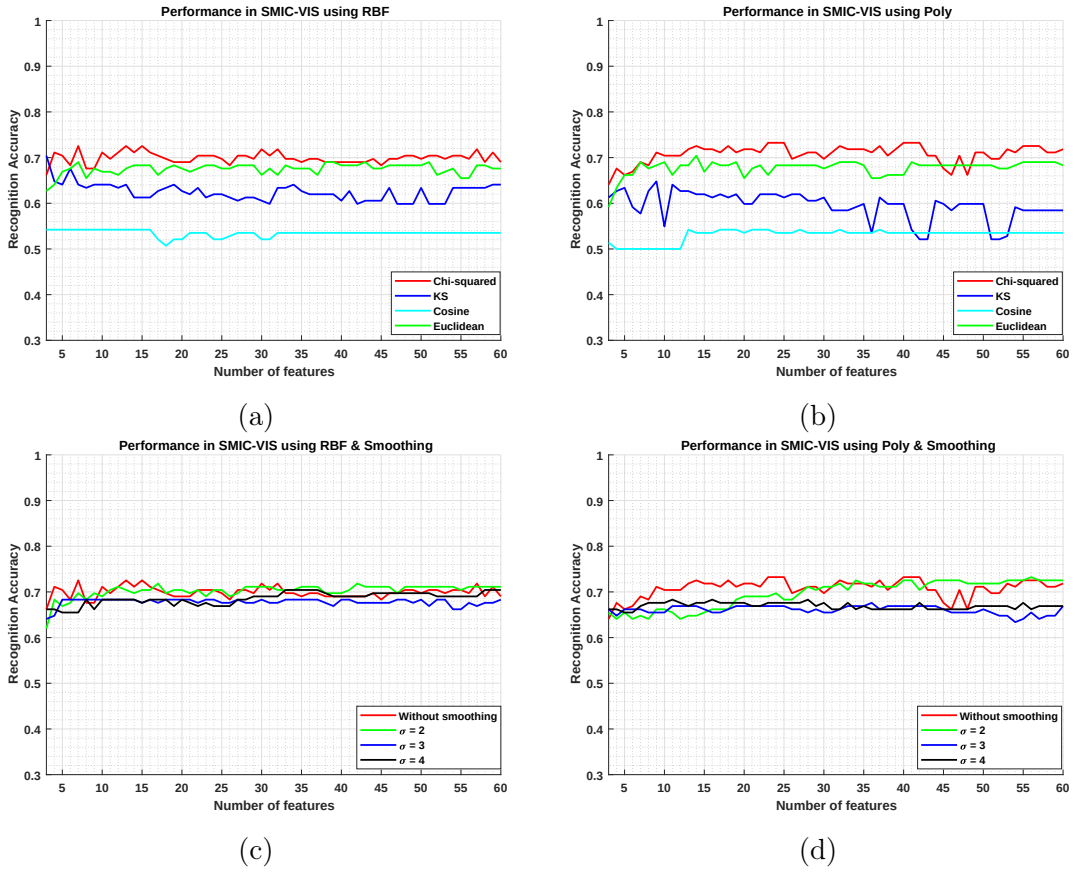


FIGURE 7: The results in the SMIC-VIS dataset. **(a)** Comparison of four difference metrics regarding the influence of feature length on the recognition performance using the RBF kernel. **(b)** Comparison of four difference metrics regarding the influence of feature length on the recognition performance using the polynomial kernel. **(c)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the RBF kernel. **(d)** Comparison of standard deviation of the Gaussian filter regarding the influence of feature length on the recognition performance using the polynomial kernel.

shows a summary of the mean and the maximum recognition rate for each combination of method. Combining the polynomial kernel and chi-squared distance yields the highest mean accuracy 70.80%. The combination of the chi-squared distance and the polynomial kernel with a number of feature $23-25$, $40-42$ achieves the best result 73.24%.

Table 4: Recognition rate comparisons between the RBF and polynomial kernel as well as comparisons between four dissimilarity metrics in VIS dataset

Method	RBF				Poly			
	CS	KS	COS	EUC	CS	KS	COS	EUC
Mean	69.94%	62.47%	53.46%	67.51%	**70.80%**	59.52%	53.06%	67.78%
Max	72.54%	70.42%	54.23%	69.01%	**73.24%**	64.79%	54.23%	70.42%

From Fig. 7 (c) and (d), applying the Gaussian filter does not gain any improvements.

Conclusion

This appendix investigate four dissimilarity metrics and the Gaussian smoothing for micro-expressions detection. Since the chi-squared distance outperforms other three dissimilarity metrics in most cases, it is selected as the dissimilarity metric in Chapter 4 and 5. For the database with more samples, exploiting the Gaussian smoothing with $\sigma = 2$ is a good choice. For these small databases, it is not necessary to apply the Gaussian smoothing.

Publication

1. Hua Lu, Kidiyo Kpalma and Joseph Ronsin, Micro-expression motion detection using integral projections, Journal of WSCG. Vol.25, 2017. No.2

Bibliography

[AHP06] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns : Application to face recognition. IEEE transactions on pattern analysis and machine intelligence, 28(12) :2037–2041, 2006.

[BA96] Michael J Black and Paul Anandan. The robust estimation of multiple motions : Parametric and piecewise-smooth flow fields. Computer vision and image understanding, 63(1) :75–104, 1996.

[BBPW04] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. Computer Vision-ECCV 2004, pages 25–36, 2004.

[Bet12] Vinay Bettadapura. Face expression recognition and analysis : the state of the art. arXiv preprint arXiv :1203.6722, 2012.

[BGV92] Bernhard E Boser, Isabelle Guyon, and Vladimir Vapnik. A Training Algorithm for Optimal Margin Classifiers. COLT, pages 144–152, 1992.

[BP93] Roberto Brunelli and Tomaso Poggio. Face recognition : Features versus templates. IEEE transactions on pattern analysis and machine intelligence, 15(10) :1042–1052, 1993.

[Bro91] Donald E Brown. Human universals. McGraw-Hill New York, 1991.

[BSL+11] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. International Journal of Computer Vision, 92(1) :1–31, 2011.

[Bur98] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2) :121–167, 1998.

[CAE07] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. The handbook of emotion elicitation and assessment, pages 203–221, 2007.

[CH92] John G Carlson and Elaine Hatfield. Psychology of emotion. Harcourt Brace Jovanovich, 1992.

[CKM07] Chi-Ho Chan, Josef Kittler, and Kieron Messer. Multi-scale local binary pattern histograms for face recognition. Advances in biometrics, pages 809–818, 2007.

[CL11] Chih-Chung Chang and Chih-Jen Lin. Libsvm : a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3) :27, 2011.

[CRHV09] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1932–1939. IEEE, 2009.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine learning, 20(3) :273–297, 1995.

[Dan10] Benjamin Smith Daniel. Photos : Dr. cal lightmanâĂŹs seven universal micro-expressions. 2010.

[DCX+15] F De, WS Chu, X Xiong, F Vicente, et al. Intraface. In AFGR, 2015.

[DFKS11] Axel Dreves, Francisco Facchinei, Christian Kanzow, and Simone Sagratella. On the solution of the kkt conditions of generalized nash equilibrium problems. SIAM Journal on Optimization, 21(3) :1082–1108, 2011.

[DK98] Bella M DePaulo and Deborah A Kashy. Everyday lies in close and casual relationships. Journal of personality and social psychology, 74(1) :63, 1998.

[DKK+96] Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. Journal of personality and social psychology, 70(5) :979, 1996.

[DLC+16] AK Davison, Cliff Lansley, Nicolas Costen, Kevin Tan, and Moi Hoon Yap. Samm : micro-facial movement dataset. IEEE Transactions on Affective Computing, 2016.

[DP98] Charles Darwin and Phillip Prodger. The expression of the emotions in man and animals. Oxford University Press, USA, 1998.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005.

[DTS06] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In European conference on computer vision, pages 428–441. Springer, 2006.

[DYC+14] Adrian K Davison, Moi Hoon Yap, Nicholas Costen, Kevin Tan, Cliff Lansley, and Daniel Leightley. Micro-facial movements : an investigation on spatio-temporal descriptors. In European Conference on Computer Vision, pages 111–123. Springer, 2014.

[DYL15] Adrian K Davison, Moi Hoon Yap, and Cliff Lansley. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on, pages 1864–1869. IEEE, 2015.

[EF69] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. Psychiatry, 32(1) :88–106, 1969.

[EF74] Paul Ekman and Wallace V Friesen. Detecting deception from the body or face. Journal of personality and Social Psychology, 29(3) :288, 1974.

[EF75] Paul Ekman and Wallace V Friesen. Unmasking the face : A guide to recognizing emotions from facial cues, 1975.

[EF76] Paul Ekman and Wallace V Friesen. Measuring facial movement. Environmental psychology and nonverbal behavior, 1(1) :56–75, 1976.

[EF77] Paul Ekman and Wallace V Friesen. Facial action coding system. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[EF86] Paul Ekman and Wallace V Friesen. A new pan-cultural facial expression of emotion. Motivation and emotion, 10(2) :159–168, 1986.

[EFO+87] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. Journal of personality and social psychology, 53(4) :712, 1987.

[Ekm02] P Ekman. Microexpression training tool(mett). San Francisco University of California, 2002.

[Ekm03] P Ekman. Mett. micro expression training tool. CD-ROM. Oakland, 2003.

[Ekm07] Paul Ekman. Emotions revealed : Recognizing faces and feelings to improve communication and emotional life. Macmillan, 2007.

[Ekm09] Paul Ekman. Telling lies : Clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company, 2009.

[EL09a] Jennifer Endres and Anita Laidlaw. Micro-expression recognition training in medical students : a pilot study. BMC Medical Education, 9(1) :47, Jul 2009.

[EL09b] Jennifer Endres and Anita Laidlaw. Micro-expression recognition training in medical students : a pilot study. BMC medical education, 9(1) :47, 2009.

[ER97] Paul Ekman and Erika L Rosenberg. What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

[GLC00] Guodong Guo, Stan Z Li, and Kapluk Chan. Face recognition by support vector machines. In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pages 196–201. IEEE, 2000.

[GZZ10] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. IEEE Transactions on Image Processing, 19(6) :1657–1663, 2010.

[HF11] Carolyn M Hurley and Mark G Frank. Executing facial control during deception situations. Journal of Nonverbal Behavior, 35(2) :119–131, 2011.

[HI66a] Ernest A Haggard and Kenneth S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In Methods of research in psychotherapy, pages 154–165. Springer, 1966.

[HI66b] Ernest A Haggard and Kenneth S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In Methods of research in psychotherapy, pages 154–165. Springer, 1966.

[HPA04] Abdenour Hadid, Matti Pietikainen, and Timo Ahonen. A discriminative feature space for detecting and recognizing faces. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–II. IEEE, 2004.

[HPS06] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with center-symmetric local binary patterns. In ICVGIP, volume 6, pages 58–69. Springer, 2006.

[HS81] Berthold KP Horn and Brian G Schunck. Determining optical flow. Artificial intelligence, 17(1-3) :185–203, 1981.

[HWL+16] Xiaohua Huang, Sujing Wang, Xin Liu, Guoying Zhao, Xiaoyi Feng, and Matti Pietikainen. Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection. arXiv preprint arXiv :1608.02255, 2016.

[HWZP15] Xiaohua Huang, Su-Jing Wang, Guoying Zhao, and Matti Piteikainen. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 1–9, 2015.

[HZH+16a] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing, 175 :564–578, 2016.

[HZH+16b] Xiaohua Huang, Guoying Zhao, Xiaopeng Hong, Wenming Zheng, and Matti Pietikäinen. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing, 175 :564–578, 2016.

[JBY+17] Xitong Jia, Xianye Ben, Hui Yuan, Kidiyo Kpalma, and Weixiao Meng. Macro-to-micro transformation model for micro-expression recognition. Journal of Computational Science, 2017.

[KCT00] T. Kanade, J. F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In Automatic Face and Gesture Recognition,

2000. Proceedings. Fourth IEEE International Conference on, pages 46–53. IEEE, 2000.

[KSBM01] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 13(3) :637–649, 2001.

[LaB47] Weston LaBarre. The cultural basis of emotions and gestures. Journal of personality, 16(1) :49–68, 1947.

[LCK+10] Patrick Lucey, Jeffrey 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. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 94–101. IEEE, 2010.

[Lee96] Tai Sing Lee. Image representation using 2d gabor wavelets. IEEE Transactions on pattern analysis and machine intelligence, 18(10) :959–971, 1996.

[LHM+15] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Reading hidden emotions : spontaneous micro-expression spotting and recognition. arXiv preprint arXiv :1511.00423, 2015.

[LHM+17] X. Li, X. HONG, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen. Towards reading hidden emotions : A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Transactions on Affective Computing, PP(99) :1–1, 2017.

[LKR17] Hua Lu, Kidiyo Kpalma, and Joseph Ronsin. Micro-expression detection using integral projections. Journal of WSCG, 25(2) :87–96, 2017.

[LNSP17] Anh Cat Le Ngo, John See, and C-W Raphael Phan. Sparsity in dynamics of spontaneous subtle emotion : Analysis & application. IEEE Transactions on Affective Computing, 2017.

[LPH+13] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. A spontaneous micro-expression database : Inducement, collection and baseline. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–6, April 2013.

[LSPW16] Sze-Teng Liong, John See, Raphael Chung-Wei Phan, and KokSheik Wong. Less is more : Micro-expression recognition from video using apex frame. arXiv preprint arXiv :1606.01721, 2016.

[LZY+16] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing, 7(4) :299–310, 2016.

[MF92] Batja Mesquita and Nico H Frijda. Cultural variations in emotions : a review. Psychological bulletin, 112(2) :179, 1992.

[MH11] David Matsumoto and Hyi Sung Hwang. Evidence for training the ability to read microexpressions of emotion. Motivation and Emotion, 35(2) :181–191, 2011.

[MLWC+00] David Matsumoto, Jeff LeRoux, Carinda Wilson-Cohn, Jake Raroque, Kristie Kooken, Paul Ekman, Nathan Yrizarry, Sherry Loewinger, Hideko Uchida, Albert Yee, et al. A new test to measure emotion recognition ability : Matsumoto and ekman's japanese and caucasian brief affect recognition test (jacbart). Journal of Nonverbal Behavior, 24(3) :179–209, 2000.

[MM01] O. L. Mangasarian and David R. Musicant. Lagrangian Support Vector Machines. Journal of Machine Learning Research, 2001.

[MZP14a] Antti Moilanen, Guoying Zhao, and Matti Pietikäinen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In Pattern Recognition (ICPR), 2014 22nd International Conference on, pages 1722–1727. IEEE, 2014.

[MZP14b] Antti Moilanen, Guoying Zhao, and Matti Pietikainen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In 2014 22nd International Conference on Pattern Recognition (ICPR), pages 1722–1727. IEEE, 2014.

[NE86] Hans-Hellmut Nagel and Wilfried Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5 :565–593, 1986.

[OFG97] Edgar Osuna, Robert Freund, and Federico Girosit. Training support vector machines : an application to face detection. In Computer vision

and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on, pages 130–136. IEEE, 1997.

[OGL51] K. N. OGLE. The perception of the visual world. Science, 113, 05 1951.

[OPH96] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 29(1) :51–59, 1996.

[OPM00] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In European Conference on Computer Vision, pages 404–420. Springer, 2000.

[OPM02] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7) :971–987, 2002.

[OT90] Andrew Ortony and Terence J Turner. What's basic about basic emotions ? Psychological review, 97(3) :315, 1990.

[OVOP01] Timo Ojala, Kimmo Valkealahti, Erkki Oja, and Matti Pietikäinen. Texture discrimination with multidimensional distributions of signed gray-level differences. Pattern Recognition, 34(3) :727–739, 2001.

[PKO09] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In Crime Detection and Prevention (ICDP 2009), 3rd International Conference on, pages 1–6. IET, 2009.

[Pla98] John Platt. Sequential minimal optimization : A fast algorithm for training support vector machines. 1998.

[PLZP11] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 1449–1456. IEEE, 2011.

[Pre13] Peter M Prendergast. Anatomy of the face and neck. In Cosmetic Surgery, pages 29–45. Springer, 2013.

[PTB08] Stephen Porter and Leanne Ten Brinke. Reading between the lies : Identifying concealed and falsified emotions in universal facial expressions. Psychological science, 19(5) :508–514, 2008.

[PW08] Frederic I Parke and Keith Waters. Computer facial animation. CRC Press, 2008.

[QWY+17] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. Cas (me)ˆ 2 : A database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Transactions on Affective Computing, 2017.

[QWYF16] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, and Xiaolan Fu. Cas (me) 2 : A database of spontaneous macro-expressions and micro-expressions. In International Conference on Human-Computer Interaction, pages 48–59. Springer, 2016.

[RHP13] John A Ruiz-Hernandez and Matti Pietikäinen. Encoding local binary patterns using the re-parametrization of the second order gaussian jet. In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pages 1–6. IEEE, 2013.

[RT01] Roman Rosipal and Leonard J Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. Journal of machine learning research, 2(Dec) :97–123, 2001.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 40(2) :99–121, 2000.

[Rus94] James A Russell. Is there universal recognition of emotion from facial expression ? a review of the cross-cultural studies. Psychological bulletin, 115(1) :102, 1994.

[SB91] Michael J Swain and Dana H Ballard. Color indexing. International journal of computer vision, 7(1) :11–32, 1991.

[SB09] Steven M Seitz and Simon Baker. Filter flow. In Computer Vision, 2009 IEEE 12th International Conference on, pages 143–150. IEEE, 2009.

[SBF+14] Matthew Shreve, Jesse Brizzi, Sergiy Fefilatyev, Timur Luguev, Dmitry Goldgof, and Sudeep Sarkar. Automatic expression spotting in videos. Image and Vision Computing, 32(8) :476–486, 2014.

[Sch71] Martin Schechter. Principles of functional analysis, volume 2. Academic press New York, 1971.

[Sch85] Klaus Rainer Scherer. Handbook of methods in nonverbal behavior research. Cambridge University Press, 1985.

[SGGS11] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar. Macro-and micro-expression spotting in long videos using spatio-temporal strain. In 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pages 51–56, 2011.

[SGM09] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns : A comprehensive study. Image and Vision Computing, 27(6) :803–816, 2009.

[SKBM00] Shirish K Shevade, S Sathiya Keerthi, Chiranjib Bhattacharyya, and K R K Murthy. Improvements to the SMO algorithm for SVM regression. IEEE Trans. Neural Netw. Learning Syst., 11(5) :1188–1193, 2000.

[SRB10] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2432–2439. IEEE, 2010.

[SRB14] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. International Journal of Computer Vision, 106(2) :115–137, 2014.

[SS02] Bernhard Schölkopf and Alexander J Smola. Learning with kernels : support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[SV99] J A K Suykens and J Vandewalle. Least Squares Support Vector Machine Classifiers. Neural Processing Letters, 9(3) :293–300, June 1999.

[SWF12] Xun-bing Shen, Qi Wu, and Xiao-lan Fu. Effects of the duration of expressions on the recognition of microexpressions. Journal of Zhejiang University-Science B, 13(3) :221–230, 2012.

[Tom84] Silvan S Tomkins. Affect theory. Approaches to emotion, 163(163–195), 1984.

[TT10] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE transactions on image processing, 19(6) :1635–1650, 2010.

[TTMM00] Mäenpää Topi, Ojala Timo, Pietikäinen Matti, and Soriano Maricor. Robust texture classification by subsets of local binary patterns. In Pattern

Recognition, 2000. Proceedings. 15th International Conference on, volume 3, pages 935–938. IEEE, 2000.

[Vap63] Vladimir Vapnik. Pattern recognition using generalized portrait method. Automation and remote control, 24 :774–780, 1963.

[VDFS09] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9 : A database of political debates for analysis of social interactions. In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pages 1–4. IEEE, 2009.

[W+99] Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. Advances in Kernel Methods-Support Vector Learning, 6 :69–87, 1999.

[Wei10] Sharon Weinberger. Intent to deceive ? Nature, 465(7297) :412, 2010.

[WJYWZ+14] X.-B. Li W.-J. Yan, S.-J. Wang, G.-Y. Zhao, Y.-J. Liu, Y.-H. Chen, and X.-L. Fu. Casme ii : An improved spontaneous micro-expression database and the baseline evaluation. PLOS ONE, 9 :1–8, 2014.

[WKSL13] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision, 103(1) :60–79, 2013.

[WM16] Kasia Wezowski and Dominika Maison. Reading facial expression to understand human emotions : Micro-expressions training videos (metv) : The new tool for experimental economics. In Selected Issues in Experimental Economics, pages 135–149. Springer, 2016.

[WPZ+09] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. An improved algorithm for tv-l1 optical flow. In Statistical and geometrical approaches to visual motion analysis, pages 23–45. Springer, 2009.

[WRS+12] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. ACM Transactions on Graphics, 31(4) :1–8, 2012.

[WSB09] Gemma Warren, Elizabeth Schertler, and Peter Bull. Detecting deception from emotional and unemotional cues. Journal of Nonverbal Behavior, 33(1) :59–69, 2009.

[WSF11] Qi Wu, Xunbing Shen, and Xiaolan Fu. The machine knows what you are hiding : an automatic micro-expression recognition system. Affective Computing and Intelligent Interaction, pages 152–162, 2011.

[WSPO14] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. Lbp with six intersection points : Reducing redundant information in lbp-top for micro-expression recognition. In Asian Conference on Computer Vision, pages 525–537. Springer, 2014.

[WW12] K Wezowski and P Wezowski. The micro expressions book for business. New Vision, Antwerp, 127, 2012.

[WWQ+17] Su-Jing Wang, Shuhang Wu, Xingsheng Qian, Jingxiu Li, and Xiaolan Fu. A main directional maximal difference analysis for spotting facial movements from long-term videos. Neurocomputing, 230 :382–389, 2017.

[WYL+14] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In Pattern Recognition (ICPR), 2014 22nd International Conference on, pages 4678–4683. IEEE, 2014.

[WYL+15] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. IEEE Transactions on Image Processing, 24(12) :6034–6047, 2015.

[WYS+16] Su-Jing Wang, Wen-Jing Yan, Tingkai Sun, Guoying Zhao, and Xiaolan Fu. Sparse tensor canonical correlation analysis for micro-expression recognition. Neurocomputing, 214 :218–232, 2016.

[WYZ+14] Su-Jing Wang, Wen-Jing Yan, Guoying Zhao, Xiaolan Fu, and Chun-Guang Zhou. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In Workshop at the European Conference on Computer Vision, pages 325–338. Springer, Cham, 2014.

[XFP+16] Zhaoqiang Xia, Xiaoyi Feng, Jinye Peng, Xianlin Peng, and Guoying Zhao. Spontaneous micro-expression spotting via geometric deformation modeling. Computer Vision and Image Understanding, 147 :87–94, 2016.

[XT13] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR 2013), pages 532–539, 2013.

[XZW17] Feng Xu, Junping Zhang, and James Z Wang. Microexpression identification and categorization using a facial dynamics map. IEEE Transactions on Affective Computing, 8(2) :254–267, 2017.

[YWL+13a] Wen-Jing Yan, Q. Wu, Yong-Jin Liu, Su-Jing Wang, and X. Fu. Casme database : A dataset of spontaneous micro-expressions collected from neutralized faces. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–7, April 2013.

[YWL+13b] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. How fast are the leaked facial expressions : The duration of micro-expressions. Journal of Nonverbal Behavior, 37(4) :217–230, 2013.

[ZG00] Victor W Zue and James R Glass. Conversational interfaces : Advances and challenges. Proceedings of the IEEE, 88(8) :1166–1180, 2000.

[ZL03] Dengsheng Zhang and Guojun Lu. Evaluation of similarity measurement for image retrieval. In Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on, volume 2, pages 928–931. IEEE, 2003.

[ZP07] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 29(6) :915–928, 2007.

[ZR16] Elham Zarezadeh and Mehdi Rezaeian. Micro expression recognition using the eulerian video magnification method. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 7(3) :43–54, 2016.

List of figures

List of tables

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Video Analysis for Micro-Expression Spotting and Recognition

Nom Prénom de l'auteur : LU HUA

Membres du jury :
- Monsieur KPALMA Kidiyo
- Monsieur RONSIN Joseph
- Monsieur BOURDON Pascal
- Monsieur ALATA Olivier
- Monsieur ROSENBERGER Christophe
- Monsieur DAOUDI Mohamed
- Madame SOLADIE Catherine

Président du jury : ROSENBERGER christophe

Date de la soutenance : 05 Avril 2018

Reproduction de la these soutenue

☑ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 05 Avril 2018

Signature du président de jury

Le Directeur,

M'hamed DRISSI

INSTITUT NATIONAL des SCIENCES APPLIQUÉES - RENNES

Résumé

Les principales contributions de cette thèse, en analyse d'image, portent sur l'étude des caractéristiques de repérage et de reconnaissance des micro-expressions. Les approches d'analyse d'expressions faciales dans le domaine de la vision par ordinateur consistent à les détecter et à les classer dans des vidéos. Par rapport à la macro-expression, une microexpression induit dans une partie du visage un changement rapide durant moins d'une demi-seconde. De plus, cette subtile apparition dans une partie du visage rend difficile sa détection et sa reconnaissance. Ces dernières années ont connu un intérêt croissant pour des algorithmes d'extraction automatique de micro-expressions faciales. Cela a été motivé par des applications dans des contextes à enjeux élevés tels les enquêtes criminelles, les points de contrôle des aéroports et des transports en commun, le contre-terrorisme, Le choix de caractéristiques faciales efficaces joue un rôle crucial dans l'analyse des micro-expressions.

Ce travail se concentre sur la partie d'extraction de caractéristiques, en proposant diverses méthodes pour les tâches de détection et de reconnaissance de micro-expression. La détection constitue la première étape dans l'analyse des micro-expressions. Les méthodes de détection existantes basées sur des caractéristiques, tels les motifs binaires locaux, l'histogramme de gradients orientés, le flux optique, souffrent de complexité de mise en oeuvre entraînant un problème d'implémentation en temps réel. Ainsi, dans cette thèse, une méthode de détection basée sur la projection intégrale est proposée pour résoudre ce problème. Cependant, toutes les caractéristiques ci-dessus sont extraites des visages recadrés et rognées ; ce qui cause, généralement, un décalage résiduel entre les images. Pour résoudre ce problème, est proposée une autre méthode de détection basée sur des caractéristiques géométriques. Cette dernière exploite les distances géométriques entre des points clés du visage sans nécessité de recadrer l'image. Ceci permet de capturer des déplacements géométriques subtils le long des séquences et s'avère approprié pour différentes tâches d'analyse faciale qui requièrent une grande vitesse de calcul.

Parmi les caractéristiques de reconnaissance de micro expressions existantes, celles de mouvement basées sur le flux optique présentent des avantages dans la caractérisation de mouvements subtils sur le visage. Toutefois, il reste difficile de déterminer les emplacements précis de chaque mappage de traits du visage entre les différentes trames par flux optique, même si les images ont été alignées. Un tel problème peut donner lieu à une mauvaise estimation, à la fois, de l'orientation et de l'amplitude associées au flux optique. Pour y pallier, nous proposons une nouvelle approche basée sur les histogrammes de frontière de mouvement. Elle permet de supprimer les mouvements inattendus causés par un mauvais recalage résiduel apparaissant entre les images recadrées tout en capturant le mouvement relatif caractérisant la microexpression. Cette caractéristique est générée en combinant les composantes horizontales et verticales du différentiel de flux optique. Les différents développements de ce travail ont conduit à des études comparatives avec des approches de l'état de l'art sur des bases de données bien connues et exploitées par la communauté du domaine. Les résultats expérimentaux, ainsi obtenues, montrent l'efficacité de nos contributions.

Abstract

Recent years, there has been an increasing interest in the computer vision in automatic facial micro-expression algorithms. This has been driven by applications in high-stakes contexts such as criminal investigations, airport and mass transit checkpoints, counter terrorism, and so on. Micro-expression approaches in computer vision area consist of detecting and classifying them from videos. Compared to macro-expression, a micro-expression involves a rapid change which lasts less than a half of second, and moreover, its subtle appearance in part of the face makes detection and recognition difficult to achieve. Effective facial features play a crucial role for micro-expression analysis. This thesis focuses on the feature extraction parts, by developing various feature extraction methods for types of micro-expression detection and recognition tasks.

The detection of micro-expressions is the first step for its analysis. This thesis aims to spot micro-expressions from videos. Existing detection methods based on features, such as the local binary patterns, the histogram of gradient and the optical flow suffer difficulties in computation consuming leading to real-time implementation problem. Thus, in this thesis, the spotting method based on integral projection is exploited to address this problem. However, all the above features are extracted from cropped faces which usually cause residual misregistration that appears between images. In order to deal with this issue, another detection method based on geometrical feature is proposed. It involves the geometrical distances between facial key-points without the need of cropping face. This captures subtle geometric displacements along sequences and is proved to be suitable for different facial analysis tasks that require high computational speed.

For micro-expression recognition, motion features based on the optical flow have advantages in characterizing subtle movements on face among the existing recognition features. It is still a difficult problem for optical flow to determine the accurate locations of each facial feature mappings between different images even though the face images have been aligned. Such an issue may give rise to wrong orientation and magnitude estimation associated to the optical flow field. In order to address this problem, the motion boundary histograms are considered. It can remove unexpected motions caused by residual mis-registration that appears between images cropped from different frames. Nevertheless, the relative motion can be captured. Based on the motion boundary, a new descriptor the Fusion Motion Boundary Histograms is introduced. This feature is generated by combing both the horizontal and the vertical components of the differential of optical flow as inspired from the motion boundary histograms.

The main contributions of this thesis lie at the study of features for micro-expressions spotting and recognition. Experiments on the micro-expression databases show the effectiveness of the presented contributions.

INSA RENNES

N° d'ordre : 18 ISAR 08/D18.08

Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coëmes • CS 70839 • F-35708 Rennes Cedex 7
Tel : 02 23 23 82 00 - Fax : 02 23 23 83 96