Proposition de recherche doctorale

Détection et indexation de visages parlants dans des séquences vidéo

Mots clés :

- **Directeur de thèse** : Gérard CHOLLET
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire Traitement et Communication de l'Information
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Avec le développement rapide de moyens de communication audiovisuels et la mise en ligne de séquences vidéo sur internet, nous constatons une prolifération des contenus multimédia. Il devient indispensable de développer des méthodes pour faciliter l'accès à ces contenus. La détection et reconnaissance de personnes dans les vidéos est une clé d'indexation pertinente pour la recherche d'information. L'indexation audio-visuelle des personnes a pour but de permettre à un utilisateur de localiser les interventions de certaines personnalités dans des séquences vidéo. L'objectif de la thèse est de proposer un algorithme robuste d'indexation des personnes basé sur les deux informations audio et visuelle. Les systèmes d'indexation basés sur le locuteur sont sensibles à l'environnement acoustiques et la complexité du scénario (discours spontané, stress, grippes ?). Les systèmes d'indexation basés uniquement sur la modalité visage sont sensibles aux conditions d'éclairage, aux variations de pose et d'expressions faciales. Les sensibilités étant différentes dans chacune des modalités, la fusion est potentiellement plus robuste aux dégradations de chacune des modalités prises séparément. Malheureusement, la fusion des deux modalités présente certaines difficultés due au fait que les plans visuels d'une personne ne sont pas souvent synchronisés avec les séquences de ses interventions audio (ie une personne peut parler et ne pas être filmée ou le contraire). Dans ce cas, la synchronisation du mouvement des lèvres avec le signal de parole peut être un indicateur pour assurer la bonne association les modalités. Pour tenter de résoudre ces problèmes, l?utilisation de modèle générique 3D de visage parlant sera développée et adapté à chaque individu rencontré. Ce modèle générique peut être utilisé de manières différentes. Il peut être utile pour générer des données synthétiques annotées nécessaires pour différentes méthodes statistiques à base d?apprentissage à partir d?exemples. Ce modèle 3D permettra aussi de s'adapter aux variations de pose observées dans les séquences vidéo. Nous nous intéresserons particulièrement aux diverses stratégies de fusion, notamment la fusion de caractéristiques, sujet de recherche émergent, dont le butsera de créer un nouveau modèle de représentation, mixant des caractéristiques bas niveaux des signaux images et son. Le signal audiovisuel sera segmenté en séquences de trames dans lesquelles la même personne s'exprime. Ces segments seront regroupés par identité. L'évaluation sera effectuée sur des données publiques en suivant des protocoles permettant la reproductibilité des résultats.