Regroupement d'Activités de la Main Non-étiquetées

Yasser BOUTALEB[1,2], Catherine SOLADIE[2], Nam-Duong DUONG[1], Amine KACETE[1], Jerome ROYAN[1], Renaud SEGUIER[2]

[1] IRT b-com
1219 Avenue des Champs Blancs, 35510 Cesson-Sevigne, France

[2] IETR/CentraleSupelec
Avenue de la Boulaie, 35510 Cesson-Sevigne, France

{yasser.Boutaleb, nam-duong.duong, amine.kacete, jerome.royan}@b-com.fr,
{catherine.soladie, renaud.seguier}@centralesupelec.fr

Résumé – Nous proposons, dans cet article, une nouvelle approche qui utilise l'adaptation de domaine non supervisée (UDA) pour le regroupement d'activités de la main non étiquetées, c'est à dire le regroupement d'activités similaires dans des groupes distincts. L'approche proposée vise à exploiter les connaissances tirées des échantillons étiquetés du domaine source pour catégoriser les échantillons non étiquetés du domaine cible. À cette fin, nous introduisons une nouvelle fonction de perte, basée sur l'apprentissage par métrique, pour apprendre une représentation hautement discriminante, tout en préservant une bonne précision de reconnaissance des activités sur le domaine source. La représentation apprise est utilisée comme un espace de dimension réduite pour regrouper automatiquement les échantillons non étiquetés. En outre, pour garantir de meilleurs résultats de regroupement, nous avons proposé une stratégie statistique et consensuelle de regroupement. L'approche proposée est évaluée sur le base de données FPHA.

Abstract – We propose in this paper a new approach based on unsupervised domain adaptation (UDA) for 3D skeleton hand activity clustering. It aims at exploiting the knowledge-driven from labeled samples of the source domain to categorize the unlabeled ones of the target domain. To this end, we introduce a novel metric learning-based loss function to learn a highly discriminative representation while preserving a good activity recognition accuracy on the source domain. The learned representation is used as a low-level manifold to cluster unlabeled samples. In addition, to ensure the best clustering results, we proposed a statistical and consensus-clustering-based strategy. The proposed approach is experimented on the real-world FPHA data set.

1 Introduction

La compréhension de l'activité de la main à la première personne est un sujet d'actualité dans le domaine de la vision par ordinateur, avec de divers applications, telles que l'interaction homme-machine, la robotique humanoïde, et la réalité virtuelle/augmentée.

Les données de squelette 3D sont l'une des modalités les plus efficaces pour la reconnaissance de l'activité humaine. En effet, elles fournissent une description robuste qui dépasse les problèmes courants d'imagerie RVB, tels que la soustraction du fond et la variation de la lumière. À cette fin, plusieurs approches de reconnaissance d'activité basées sur le squelette 3D ont été proposées. La plupart d'entre elles sont basées sur des architectures supervisées de réseaux neuronaux (NN), qui se sont avérées efficaces lorsqu'une grande quantité de données est disponible. Cependant, dans les applications de reconnaissance d'activité du monde-réel, les échantillons de squelette 3D étiquetés sont généralement difficiles ou coûteux à obtenir car ils nécessitent l'effort d'annotateurs experts. Cela a motivé les chercheurs à exploiter des échantillons non étiquetés en appliquant un apprentissage entièrement non-supervisé pour les regrouper en plusieurs catégories [1]. Cependant, le regroupement des activités de la main 3D sans connaissances préalables reste un problème difficile. Cela est dû à la forte variation intra-classe et à la similarité inter-classe que présentent les activités de la main.

À cet effet, nous proposons une approche de catégorisation (regroupement) des activités de la main non-étiquetés basée sur le squelette 3D, qui s'inscrit dans le paradigme d'UDA. Tout d'abord, nous pré-entraînons un NN sur des échantillons étiquetés du domaine source. Ensuite, nous essayons de résoudre l'UDA en utilisant l'espace des caractéristiques du modèle pré-entraîné en tant qu'espace de dimension réduite, pour catégoriser les échantillons non-étiquetés du domaine cible, en se basant sur des méthodes classiques de regroupement. Le modèle pré-entraîné, doit projeter les activités dans un espace de caractéristiques hautement discriminant (Figure 1). Cette exigence est largement discutée dans le domaine de la reconnaissance des visages [3–5]. Ainsi, nous avons conclu que l'espace caractéristique souhaité doit satisfaire deux objectifs principaux de l'apprentissage par métrique : (1) maximiser les distances

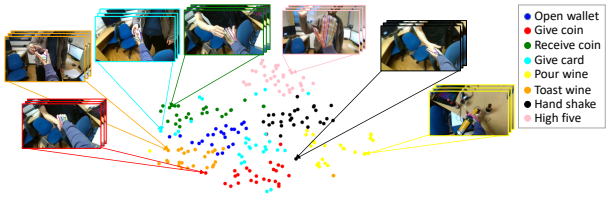

FIGURE 1 – Projection 2D des activités de la main non-étiquetées à l'aide de notre méthode, appliquée sur le scénario "social" de la base de données FPHA [2].

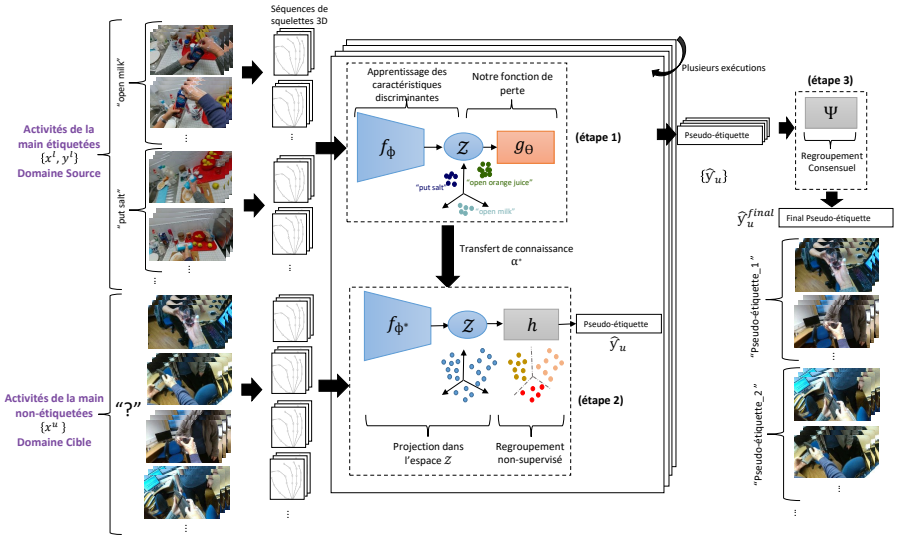

FIGURE 2 – Fonctionnement générale de la méthode proposée. Grâce à notre fonction perte on apprenons un espace de caractéristiques qui facilite le regroupement d'activités non étiquetées. De plus, notre stratégie statistique et consensuelle assure des meuliers résultats de regroupement.

inter-classes et (2) minimiser les distances intra-classes pour les activités de la main projetées. C'est l'un des principaux problèmes que nous avons abordé dans cet article, en proposant une nouvelle fonction de perte basée sur l'apprentissage par métrique. D'autre part, nous abordons le problème de la grande sensibilité des algorithmes de regroupement par rapport à l'espace de caractéristiques fourni par le modèle pré-entraîné. Pour résoudre ce problème, nous avons proposé une stratégie statistique et consusenelle (SSC).

2 Méthode proposée

Pour un ensemble donné d'échantillons d'activité de la main non-étiquetés $\{x^u\}$ d'un domaine cible, nous exploitons la connaissance présente dans les échantillons étiquetés $\{x^l,y^l\}$ du domaine source pour trouver l'ensemble inconnu correspondant de pseudo-étiquettes $\mathcal{Y}_u = \{y^u\}$. Où x^u est un échantillon particulier non-étiqueté et y^u est son étiquette correspondante que nous souhaitons retrouver. Comme l'illustre la Figure 2, la méthode se déroule en trois étapes principales que nous décrivons respectivement dans les sections 2.1, 2.2 et 2.3. Dans la section 2.4, nous détaillons notre fonction de perte proposée.

2.1 Apprentissage supervisé (étape 1)

Tout d'abord, nous pré-entraînons le réseau neuronal de reconnaissance d'activité de la main basée sur le squelette 3D, qu'on note F_W avec n le nombre couches, et qu'on définit comme suite :

$$F_W = F^n_{w_n} \circ F^{n-1}_{w_{n-1}} \circ ... \circ F^1_{w_1} \quad (1)$$

Tel que $F^n_{w_n}$ et la dernière couche de classification qui s'appuie sur notre fonction de perte, et $F^{n-1}_{w_{n-1}} \circ ... \circ F^1_{w_1}$ correspond aux couches d'apprentissage des caractéristiques spatio-temporelles. Pour la simplicité notant $f_\phi = F^{n-1}_{w_{n-1}} \circ ... \circ F^1_{w_1}$ et $g_\theta = F^n_{w_n}$ avec $\phi = \{w_1,...,w_{n-1}\}$ et $\theta = w_n$ les paramètres d'apprentissage. L'entraînement est effectué sur des activités de la main 3D étiquetées du domaine source $\{x^l,y^l\}$ où y^l est l'étiquette de l'échantillon particulier x^l. Cette opération est effectuée de manière supervisée classique, comme le montre la Figure 2 (étape 1). Le NN adopté est basé sur l'architecture proposée par [6]. Nous avons choisi cette architecture pour ses performances de reconnaissance d'activités et sa facilité de reproductibilité.

Grâce à notre fonction de perte, cette première étape aboutit à un modèle pré-entraîné $f_{\phi^*}(.)$ avec des paramètres optimisés ϕ^*, qui projette les activités dans un espace de caractéristiques très discriminant $\mathcal{Z} \subset \mathbb{R}^d$ où d est la dimension de cet espace.

2.2 Regroupement d'activités (étape 2)

Dans cette deuxième étape, nous projetons l'ensemble des séquences d'activités de la main 3D non étiquetées $\{x^u\}$ dans l'espace $\mathcal{Z}$. Pour cela, nous faisons passer ces activités par le NN pré-entraîné $f_{\phi^*}(x^u) = z^u \in \mathcal{Z}$, ce qui donne un ensemble d'activités projetées $\{z^u\}$. Ensuite, nous appliquons un regroupement classique, que nous formulons comme suit : $h(\{z^u\}) = \hat{\mathcal{Y}}_u$ où $h(.)$ est la fonction de regroupement et $\hat{\mathcal{Y}}_u = \{\hat{y}^u\}$ sont les pseudo-étiquettes prédites pour l'ensemble non étiqueté d'activités $\{x^u\}$. La Figure 2 (étape 2) illustre cette étape.

Comme nous l'avons mentionné dans l'introduction, dans nos expériences, nous avons observé une grande sensibilité des algorithmes de regroupement au caractère aléatoire présent dans le processus de pré-entraînement de l'étape 1. Cette sensibilité est directement liée à l'initialisation aléatoire des poids ϕ et θ, des couches d'exclusion (dropout) et de l'optimisateur. À cette fin, nous avons proposé notre stratégie SSC qui consiste à répéter les étapes 1 et 2 plusieurs fois. Elle vise à stabiliser la moyenne et

la variation standard. Cette étape donne en sortie un ensemble de pseudo-étiquettes prédit $\{\hat{\mathcal{Y}}_u\}$.

2.3 Regroupement par consensus (étape 3)

Nous appliquons un regroupement par consensus basé sur la méthode proposée par [7]. Ce regroupent consensuel est appliqué sur l'ensemble de pseudo-étiquettes obtenus dans l'étape précédente : $\Psi(\{\hat{\mathcal{Y}}_u\}) = \hat{\mathcal{Y}}_u^{final}$ où $\Psi(.)$ est la fonction de regroupement par consensus et $\hat{\mathcal{Y}}_u^{final}$ est l'ensemble de pseudo-étiquettes final prédit pour les séquences d'activités de la main 3D non étiquetées $\{x^u\}$. Nous notons que le regroupement consensuel est généralement appliqué sur des prédictions résultant de différents algorithmes de regroupement ou du même algorithme avec différents hyper-paramètres. Dans notre stratégie SSC, les algorithmes de regroupement et leurs hyper-paramètres restent les mêmes alors que seule la distribution des données change en fonction des répétition des étapes 1 et 2.

2.4 Notre fonction de perte discriminative

L'apprentissage par métrique vise à apprendre une fonction de similarité. Il est combiné avec les NN pour améliorer la capacité discriminante de l'espace des caractéristiques en renforçant simultanément la compacité intra-classe et la divergence inter-classe. A cet effet, nous proposons une fonction de perte qui vise à satisfaire les objectifs d'apprentissage par métrique tout en préservant les performances de classification. Contrairement aux [3–5], notre fonction de perte est basée sur la pénalité entièrement automatisée et elle ne nécessite aucun hyper-paramètre supplémentaire. Et au lieu de se concentrer sur l'angle entre les vecteurs de caractéristique z_i et les poids θ_j la dernière couche de classification $g_\theta(.)$, nous nous concentrons sur la distance euclidienne entre ces deux vecteurs. Nous la désignons par $e_{i,j}(.,.)$ et nous la formulons comme suit :

$$e_{i,j}(z_i,\theta_j) = \|z_i - \theta_j\|_2 \tag{2}$$

Sans normaliser w_j et z_i, la norme euclidienne (L_2-norm) est apprise de manière adaptative pour minimiser la perte globale. Les caractéristiques sont donc apprises dans l'espace euclidien, un choix largement reconnu dans l'état de l'art.

Pour impliquer une pénalisation qui renforce la marge euclidienne, nous interprétons la distance e_j par une distribution t-Student. Nous fixons le degré de liberté à 1, ce qui correspond à une distribution à queue lourde de Cauchy dans l'espace de caractéristiques de faible dimension $\mathcal{Z}$. Nous désignons cette probabilité de distance par $p_{i,j}(.)$ et nous la formulons comme suit :

$$p_{i,j}(e_{i,j}) = (1+e_{i,j})^{-1} \tag{3}$$

Conformément à [8], nous utilisons la distribution t-Student à un seul degré de liberté car elle possède la propriété particulière de se rapprocher d'une loi carrée inverse pour les grandes distances $e_j(\theta_j,z_i)$ dans l'espace $\mathcal{Z}$. Cela rend la représentation des probabilités conjointes quasi-invariante aux changements d'échelle de l'espace de caractéristiques pour les échantillons projetés z_i qui sont loin de la médiane (par exemple, les échantillons bruités).

Contrairement à [3–5] qui utilisent la fonction softmax pour produire le score d'affinité probabiliste, nous normalisons directement les probabilités calculées $p_{i,j}$. Ceci est justifié par le fait que notre logits cible est basé sur une interprétation des probabilités. Il peut-être justifié aussi en raison de l'incompatibilité des fonctions de pertes basées sur la marge euclidienne avec la fonction de perte Softmax telle qu'indiquée dans [3, 9]. Nous formulons la normalisation de la probabilité comme suit :

$$\hat{y}_{i,j}^l = \frac{p_{i,j}}{\sum_{j=1}^{k} p_{i,j}} \tag{4}$$

Il en résulte un score de prédiction de probabilité que nous désignons par $\hat{y}_i^l = \{\hat{y}_{i,1}^l, ..., \hat{y}_{i,k}^l\}$. Enfin, notre fonction de perte, que nous désignons par $\mathcal{L}$, est formulée en combinaison avec la perte d'entropie croisée comme suit :

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i^l log(\hat{y}_i^l) \tag{5}$$

Où y_i^l est le codage one-hot de l'étiquette du i-ième échantillon x_i. Dans la section 3.3, nous montrons quantitativement l'efficacité de la fonction de perte proposée.

3 Expériences

3.1 Base de données

Nous avons utilisé la BDD **FPHA** [2]. Il s'agit d'une BDD accessible au public pour la reconnaissance de l'activité de la main à la première personne basée sur le squelette 3D. Nous avons adopté deux configurations de données (*Config 1* et *Config 2*) : dans *Config 1*, nous avons utilisé le scénario "social" comme domaine cible et "kichen + office" comme domaine source. Et dans *Config 2*, nous avons utilisé le scénario "office" comme domain cible alors que "kichen + social" sont utilisés comme domaine source.

3.2 Protocoles d'évaluation

Nous avons adopté deux protocoles (*P1* et *P2*) pour évaluer le regroupement d'activités non étiquetées des deux configurations *Config 1* et *Config 2*. Dans *P1*, nous supposons que le nombre de groupes possibles dans les domaines cibles est inconnu, et nous souhaitent trouver. Par conséquent, nous effectuons le regroupement sans donner la vérité terrain du nombre de groupes k en utilisant la méthode de regroupement Affinity Propagation, que nous évaluons en utilisant les métriques V-measure (Vm), l'homogénéité (Vm-h) et Adjusted Mutual Information (AMI). Dans *P2*, nous avons donné le nombre de groupes $k = 8$ et $k = 13$ pour les deux configurations *Config 1* et *Config 2* respectivement, et nous avons utilisé la méthode de regroupement Agglomératif, que nous avons évaluée en utilisant la précision du regroupement non supervisé (UCA).

Nous avons également évalué la précision de la reconnaissance d'activité (apprentissage supervisé et test) sur les deux domaines sources des deux configurations *Config 1* et *Config 2*.

3.3 Comparaison avec l'état de l'art

Les tableaux 1 et 2 présentent la comparaison des résultats du regroupement. Nous avons comparé l'impact de notre fonction de perte avec quatre fonctions de l'état de l'art. Comme prévu, la fonction de perte softmax est beaucoup moins performante par rapport aux fonction pertes basées sur l'apprentissage par métrique. Nous pouvons constater que la fonction de perte que nous proposons est équivalente à celles de l'état de l'art. Sur le jeux de données *Config 1*, nous obtenons le meilleur score Vm de 71,49% en adoptant le protocole *P1* et une UCA de 75,82% en adoptant le protocole *P2*. Dans le cas du jeux de données *Config 2*, la meilleure valeur UCA diminue à 50,60 % en utilisant la fonction de perte [3], ce qui est légèrement meilleur que notre résultat.

TABLE 1 – Résultats du regroupement en adoptant le protocole *P1*.

	Config 1				*Config 2*			
	AMI	**Vm**	**Vm-h**	$\hat{k}$	**AMI**	**Vm**	**Vm-h**	$\hat{k}$
Softmax	62.36	61.55	64.41	11	35.16	45.97	43.77	17
Liu et al. [3]	66.14	69.54	77.00	13	38.62	46.39	48.33	17
Wang et al. [4]	66.33	**69.98**	**78.98**	14	**40.07**	**47.99**	**50.46**	18
Deng et al. [5]	**66.37**	69.91	77.83	14	39.63	**47.87**	**50.59**	19
Notre méthode	**68.29**	**71.49**	**79.56**	13	**40.16**	46.09	49.97	17

TABLE 2 – Résultats du regroupement en adoptant le protocole *P2*. Les valeurs min, max et moyenne sont calculées sur une population de 100 essais. Le CC fait référence aux résultats du regroupement par consensus basé sur les 100 essais suivant notre stratégie SSC.

	Config 1				*Config 2*			
	Min	**Max**	**Mean**	**CC**	**Min**	**Max**	**Mean**	**CC**
Softmax	48.34	79.62	64.22	65.87	29.51	49.69	39.37	45.18
Liu et al. [3]	55.45	76.77	67.26	73.45	34.63	49.09	41.38	**50.60**
Wang et al. [4]	52.60	79.62	67.37	**73.93**	32.53	46.08	39.59	47.59
Deng et al. [5]	54.02	79.62	67.12	71.09	32.83	46.98	39.32	47.28
Notre méthode	54.97	88.62	67.83	**75.82**	34.03	47.59	40.37	**49.39**

Le tableau 2 confirme l'avantage de notre stratégie SSC. En effet, pour 100 essais, le résultat final du regroupement consensuel est loin du minimum, au-dessus de la moyenne, et pas si loin du maximum. La stratégie SSC permet de sélectionner une bonne prédiction de regroupement parmi plusieurs essais en respectant la contrainte d'apprentissage non supervisé.

Le tableau 3 montre que notre fonction perte elle surpasse toutes les fonctions de perte basées sur l'apprentissage par métrique de plus de 6% de précision de classification d'activités, tout en restant équivalente à la fonction de perte softmax. Cela valide la principale contribution de notre fonction de perte, qui consiste à trouver un équilibre entre l'apprentissage d'un espace favorable au regroupement qui facilite le regroupement des échantillons d'activités non étiquetées projetés du domaine cible tout en conservant une bonne précision de reconnaissance pour les échantillons d'activités étiquetées du domaine source.

TABLE 3 – Les résultats de la précision moyenne de reconnaissance d'activités sur 100 essais. L'apprentissage supervisés et le teste sont appliqués sur les données des domaines source.

	Softmax	**Liu et al. [3]**	**Wang et al. [4]**	**Deng et al. [5]**	**Notre méthode**
Config 1	**94.80**	86.67	84.72	82.53	**95.46**
Config 2	**96.24**	90.30	88.88	86.62	**96.51**

4 Conclusion

Dans le cadre de ce travail, nous avons présenté une solution d'adaptation du domaine non-supervisé pour le regroupement des activités de la main non étiquetées basée sur le squelette 3D. Les expériences basées sur un ensemble de données du monde réel, montrent que l'espace de caractéristiques appris en utilisant notre fonction de perte permet un regroupement efficace des échenillions non étiquetés du domaine cible tout en gardant une bonne précision de reconnaissance d'activités dans le domaine source. Ainsi, nous avons proposé une solution au problème de la sensibilité des algorithmes de regroupement en utilisant notre stratégie statistique et consusenelle, qui a démontrée son efficacité d'assurer un meilleur résultat de regroupement.

Références

[1] Khurram Soomro and Mubarak Shah. Unsupervised action discovery and localization in videos. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 696–705, 2017.

[2] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018.

[3] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface : Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[4] H. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wenyu Liu. Cosface : Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[5] Jiankang Deng, J. Guo, and Stefanos Zafeiriou. Arcface : Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[6] Yasser Boutaleb, Catherine Soladié, Nam-Duong Duong, Amine Kacete, Jérôme Royan, and Renaud Séguier. Efficient multi-stream temporal learning and post-fusion strategy for 3d skeleton-based hand activity recognition. In *VISIGRAPP*, 2021.

[7] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3 :583–617, 2002.

[8] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 :2579–2605, 2008.

[9] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.