(التعلم باستخدام أمثلة قليلة) للتعرف على لغة الإشارة باستخدام تقنية انتشار السمات المضمنة (Embedding Propagation)

أمجد السلمي1، خولة باجبع1، حمزة لقمان1،2، عصام لعراجي3

1جامعة الملك فهد للبترول والمعادن

2مركز البحوث المشترك للذكاء الاصطناعي SDAIA وجامعة الملك فهد للبترول والمعادن، الظهران، المملكة العربية السعودية

{g20210113, g202115030, hluqman}@kfupm.edu.sa

ServiceNow3

issam.laradji@servicenow.com

الملخص:

إن لغة الإشارة هي القناة الأساسية للتواصل لمجتمع الصم وضعاف السمع. وتتكون لغة الإشارة من العديد من الإشارات المختلفة في أشكال اليد وأنماط الحركة ووضع اليدين والوجه وأجزاء الجسم. ويجعل هذا الأمر التعرف على لغة الإشارة (SLR) مجالًا صعبًا في أبحاث الرؤية الحاسوبية. وتعالج هذه الورقة مشكلة التعرف على لغة الإشارة باستخدام التعلم من أمثلة محدودة حيث يتم استخدام النماذج المدربة على فئات الإشارات المعروفة للتعرف على الإشارات غير المرئية باستخدام التعلم من أمثلة محدودة. فقط. ويتم في هذه الطريقة استخدام مشفر محول لتعلم السمات المكانية والزمانية لإيماءات الإشارة كما يتم استخدام تقنية انتشار السمات المضمنة (embedding propagation) لإسقاط هذه السمات في مساحة. ويتم بعد ذلك تطبيق طريقة التصنيفات (label propagation) لتشذيب التضمينات الناتجةوقد أظهرت النتائج التي تم الحصول عليها أن الجمع بين طريقتي انتشار التضمين وانتشار التسميات يعزز أداء نظام التعرف على لغة الإشارة (SLR) ويحقق دقة 76.6٪ وهو ما يتجاوز دقة الشبكة النموذجية التقليدية قليلة اللقطات والتي تبلغ 72.4٪.

الكلمات الرئيسية: التعرف على لغة الإشارة، ترجمة لغة الإشارة، التعلم باستخدام أمثلة قليلة

1. المقدمة

تمثل لغة الإشارة الوسيلة الرئيسية للأشخاص الصم أو ضعاف الصوت للتواصل وتبادل المعرفة والتعبير عن مشاعرهم وبناء علاقات اجتماعية مع الآخرين (1). ومع تقدم التكنولوجيا أصبح بإمكان الأشخاص الذين يعانون من ضعف السمع والصمم التواصل مع مجتمعهم بكفاءة أكبر من خلال ترجمة لغة الإشارة إلى لغات طبيعية والعكس (2).

يعد التعرف على لغة الإشارة (SLR) أحد أكبر المشكلات التي يتم تناولها في مجال الرؤية الحاسوبية [3]. فعلى الرغم من أن معظم الإشارات لها مظهر محدد بوضوح إلا أنها تختلف قليلاً عن بعضها البعض بصريًا (4؛ 5). وهكذا فإنه لكي يشكل التعرف على لغة الإشارة تقنية شاملة فإنه يتطلب تقدمًا أساسيًا في نمذجة وتحديد الأنماط المكانية الزمنية الدقيقة لحركات اليد [3]. وهناك أيضًا عوامل أخرى تؤثر على أداء مهمة التعرف على لغة الإشارة بما في ذلك الاختلافات في منظور الرؤية (6) وتطور لغات الإشارة بمرور الوقت [7]) والاختلافات الإقليمية في لغة الإشارة (8).

يمكن تصنيف تقنية التعرف على لغة الإشارة إلى نوعين معزول ومستمر. حيث تستهدف أنظمة التعرف المعزول على لغة الإشارة إشارات من مستوى الكلمة في حين تتعرف مقاربات التعرف المستمر على لغة الإشارة على جمل لغة الإشارة (9). وقد تمت دراسة تقنية التعرف المعزول على لغة الإشارة على نطاق واسع في الأدبيات مقارنةً بتقنية التعرف المستمر (2). وكانت إحدى المشكلات الرئيسية في هذه الطرق هي الحاجة إلى عدد كبير من العينات الموضحة لكل إشارة (10)

(11) (12). حيث يجب جمع عينات موضحة لجميع الإشارات في جميع اللغات ذات الاهتمام لتلبية هذه الحاجة. ويجب أن تتضمن هذه العينات إشارات يتم التعبير عنها عدة مرات من قبل أفراد في إعدادات تسجيل مختلفة. يتم التحدث بأكثر من 140 لغة إشارة على مستوى العالم جنبًا إلى جنب مع العديد من اللهجات (13). وبالتالي فإن الطلب على الأمثلة الخاضعة للإشراف يقف عائقاً أمام توسيع نطاق التعرف على لغة الإشارة. وقد حاولت بعض الحلول في الآونة الأخيرة التغلب على هذه المشكلة باستخدام التعلم قليل الأمثلة للتعرف على الإشارات غير المرئية مع عدد قليل من العينات ذات التسميات التوضيحية (14؛ 15؛ 16؛ 3). ويشكل التعلم قليل الأمثلة تقنية لتعلم التمييز بين الفئات من خلال عدد محدود من العينات أو الأمثلة ذات التسميات.

نقدم في هذه الورقة طريقة التعلم قليل الأمثلة للتعرف على لغة الإشارة وهي طريقة مصممة خصيصًا ليمكن تعميمها على الفئات غير المرئية سابقاً. وتقبل هذه الطريقة معلومات وضعية إيماءات الإشارة وتغذيها في مشفر المحول لاستخراج مجموعة من السمات التي تشفر المعلومات المكانية والزمانية. ثم ننقل هذه السمات من مساحة السمات إلى مساحة التضمين من خلال الاستفادة من تقنية انتشار التضمين (embedding propagation) و تقنية انتشار التسميات (label propagation) معاً. وقد تم تقييم هذه الطريقة المقترحة باستخدام مجموعة بيانات (WLASL-100)، وتوضح النتائج التي تم الحصول عليها فعالية الجمع بين هاتين التقنيتين في مجال التعلم قليل الأمثلة للتعرف على لغة الإشارة.

لقد تم ترتيب هذه الورقة على النحو التالي. يبدأ القسم 2 بمراجعة الأدبيات ذات الصلة. ثم نقدم في القسم 3 طريقة التعلم قليل الأمثلة للتعرف على لغة الإشارة ويتم تقديم العمل التجريبي في القسم 4. وأخيراً يتم تقديم استنتاجاتنا وعملنا المستقبلي في القسم 5.

2. الأدبيات ذات الصلة

التعرف على لغة الإشارة (SLR): لقد تم تطوير العديد من التقنيات في العقدين الماضيين للتعرف على إيماءات لغة الإشارة (1؛ 2). وتركز غالبية هذه التقنيات بشكل أساسي على تتبع والتعرف على أيدي مستخدمي الإشارة (17؛ 18؛ 19؛ 20). حيث تمثل حركة اليدين الجزء اليدوي من لغة الإشارة بينما تمثل حركات الجسم وتعبيرات الوجه الجزء غير اليدوي من لغة الإشارة. ولم يحاول التعرف على الإشارات اليدوية وغير اليدوية في وقت واحد (21؛ 22؛ 23) سوى عدد قليل من الدراسات.

كانت هناك عدة محاولات لتطوير أساليب التعرف على لغة الإشارة القائم على التعلم العميق في السنوات الأخيرة. حيث اقترح (.Camgoz et al) (24) نموذجًا قائمًا على استخدام المحول للترجمة والتعرف المستمر على لغة الإشارة. وهنا يتم تعلم المعلومات الزمنية لإشارات الجملة بطريقة موحدة باستخدام فاقد التصنيف الزمني الاتصالي أو الارتباطي (CTC). وقد اقترحت دراسة سابقة (25) محولًا تقدميًا لترجمة جمل الكلام المنفصلة إلى تسلسلات مستمرة من التعبيرات ثلاثية الأبعاد. وقد استخدم (.Tao et al) (26) في هذا العمل زيادة تعزيزية لأبجدية الإشارة الأمريكية لمعالجة التشوهات (occlusions) غير المكتملة وتقليل تأثير تغييرات منظور الرؤية. ويتم بعد ذلك إرسال الصور المعززة الناتجة إلى الشبكات العصبية الالتفافية (CNN)بسيطة. وقد تم استخدام الشبكة الشبكات العصبية الالتفافية (CNN) في دراسة أخرى (27) لجمع العديد من التشكيلات المكانية والطيفية لصور إيماءات اليد لتوفير طريقة للكشف البصري عن هجاء الأصابع (fingerspelling) في الإيماءات. وتخلق هذه الطريقة المقترحة صورًا مكانية زمنية لحركات اليد في صيغ طيف (Gabor) ثم تستخدم الشبكة العصبية التلافيفية (CNN) المحسنة لتصنيف الإيماءات في مساحة مشتركة إلى فئات مناسبة.

تم اقتراح (SAM-SLR) للتعرف على لغة الإشارة وهو إطار عمل متعدد الوسائط يتعرف على "بنية

جسم الإنسانكوسيلة لاستغلال المعلومات متعددة الوسائط (28). وقد استخدم (Huang et al) شبكة عصبية تلافيفية ثلاثية الأبعاد لتعلم الجوانب المكانية والزمانية لإيماءات الإشارة (29). كما تم استخلاص مجموعة من السمات من أيدي المؤشر لتسليط الضوء على التغييرات المهمة في حركات اليد. وتم استخدام مجموعة بيانات تتكون من 25 إشارة لتقييم النهج المقترح والإبلاغ عن الحصول على دقة بنسبة 94.2٪. وفي عمل آخر تم تطوير نظام مختلف للتعرف على أبجدية لغة الإشارة وتم الإبلاغ عن الحصول على دقة بنسبة 98.9٪ (29).

واقترح المؤلفون في (30) نموذجًا لتقنية التعرف المعزول على لغة الإشارة باستخدام صور منتجة من إطارات ملونة تغطي تاريخ الحركات. وقد تم استخدام هذه التقنية لتلخيص المعلومات المكانية الزمنية لكل إشارة. كما تم تنفيذ نموذج يقبل صور (بنظام الألوان الأحمر والأخضر والأزرق) (RGB) وتاريخ الحركة كوحدات انتباه مكاني تعتمد على الحركة جنبًا إلى جنب مع المعمارية ثلاثية الأبعاد. ومن خلال استخدام تقنية الدمج المتأخر يتم تطبيق سمات هذا النموذج بشكل مباشر على سمات النموذج ثلاثي الأبعاد. وقد حاول (.Albanie et al) (31) التعامل مع نقص بيانات لغة الإشارة ذات التسميات التوضيحية من خلال التعرف على الكلمات الرئيسية في بث تلفزيوني لمعالج. ففي 1000 ساعة من الفيديو يتم تلقائيًا ترجمة 1000 إشارة من خلال الترجمات النصية المتوافقة زمنياً بشكل ضعيف إضافة إلى تحديد الكلمات الرئيسية الهامة. وقد قدم المؤلفون في (32) إطارًا متكاملًا للتعلم متعدد الأمثلة في أفلام لغة الإشارة.

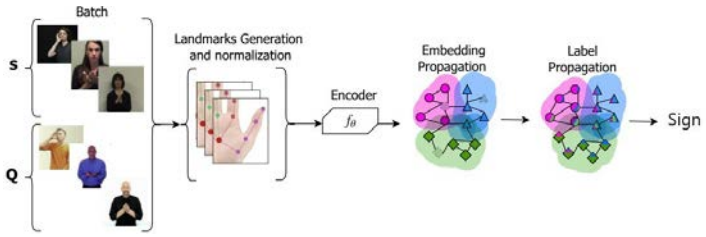
التعلم باستخدام أمثلة محدودة للتعرف على لغة الإشارة: وعلى النقيض من الطرق التقليدية التي تحتاج للإشراف للتعرف على لغة الإشارة تتعرف الطرق القائمة على التعلم قليل الأمثلة على فئات الإشارات غير المستكشفة إما من عدد قليل جدًا من عينات التدريب (أمثلة) أو بدون عينات تدريب بصرية. ويعد Cornerstone Network (CN) نموذج تعلم قليل الأمثلة اقترحه (14) ويمكنه التقليل من تأثير عينات الدعم في الظروف غير المناسبة. ويتم في هذه الشبكة استخلاص المتوسطات غير الدقيقة من عينات الدعم من عينات الإدخال واستخدامها كسمات إدخال. ويتم بعد

ذلك استخدام الشبكات العصبية مع خوارزميات التجميع لتعلم المخطط من مساحة الإدخال إلى مساحة التضمين. وكما هو الحال مع الشبكات العصبية السيامية فقد تم تدريب شبكة استخراج السمات بنفس الطريقة بحيث يتم توزيع السمات من البيانات غير المتجانسة على أوسع نطاق ممكن. وبالمثل فقد قام (Shovkoplias et al) (15) بتقصي العديد من طرق التعلم قليل الأمثلة مثل التعلم المستقل عن النموذج والتعلم الفوقي وشبكات المطابقة والشبكات النموذجية لتصنيف تسجيلات مخطط كهربية العضلات لإيماءات الصم والبكم. وقد استخدم المؤلفون في (16) متنبئًا ذا نقطة رئيسية وهو مدرب مسبقاً للاحتفاظ فقط بالمعلومات المتعلقة بالجسم واليد والوجه وتجاهل المناطق الأخرى. ويسمح هذا الأمر بمقارنة أفضل بين ضمين المتجه (Vector) حيث يتم تعلم التمثيلات الغنية من تسلسلات النقاط الرئيسية للجسم. ويتم تصنيف متجه الإدخال الجديد من خلال مقارنة مدى بعده عن بعض الأمثلة لكل فئة باستخدام خوارزمية أقرب الجيران (K-nearest neighbors)وتشابه جيب التمام والشبكات النموذجية.

Commented [DD1]: vector(تضمين المتجه)

قام (Bilge et al.) (3) بتطبيق التعلم من الصفر لتصنيف إيماءات لغة الإشارة دون أي عينات ذات تسميات توضيحية. حيث يتم إنشاء تمثيلات الفئة الدلالية من أوصاف الإشارات النصية المتوفرة بسهولة والمستمدة من قواميس لغة الإشارة. وتُستخدم هذه التمثيلات لرسم خرائط للإشارات أثناء الاستدلال على فئاتها المقابلة. وبالمثل يتم استخدام إطار التعلم من الصفر لتطوير نماذج مكانية زمنية لمناطق الجسم واليد باستخدام تمثيلات الفئة الدلالية (33). كما تم استخدام وسائط (RGB) والعمق في هذه الدراسة. وتتضمن هذه الطريقة نموذجين لمحول الرؤية يحددان أجزاء الجسم ويقسمانها إلى 9 أجزاء. ليتم بعد ذلك استخلاص مجموعة من السمات المرئية بواسطة المحول الثاني.

3. المنهجية



الشكل 1: الإطار المقترح: تم أخذ تمثيلات تقنية انتشار التضمين وتقنية انتشار التسميات من (34)

نقدم في هذا القسم نظرة عامة على مسار العملية المقترح كما هو موضح في الشكل 1. حيث تمزج هذه العملية مُشفِّر المحول (35) مع تقنية انتشار التضمين (34). ويقوم مُشفِّر المحول في البداية باستخلاص السمات من كل حركة إشارة. ليتم بعد ذلك تعيين هذه السمات للتضمينات عبر مكون انتشار التضمين. ثم نقوم بتقييم طريقتين لصقل التضمين وهما انتشار التسميات والشبكة النموذجية. ويتم أخيرًا إدخال التضمينات المكررة في مُصنِّف لتصنيف كل إشارة إلى التسمية المقابلة لها.

3.1 نموذج المحول

يتم استخدام نموذج قائم على المحول تم اقتراحه بواسطة (**35**) في عمليتنا كمستخلص للسمات لتعلم تمثيلات وضعية الجسم. ويتم استخلاص السمات باستخدام مشفر المحول بينما يتم استبدال المشفر بمكون انتشار التضمين. ويخضع كل إطار فيديو لمرحلة تقدير الوضعية قبل المعالجة ويتم تحديد معالم الرأس والجسم واليد. وبهدف تجنب مشكلة صعوبة النموذج في تعميم البيانات الجديدة وتعزيز تعميمه على مختلف البيانات يتم تعزيز البيانات الهيكلية أثناء التدريب عبر التقنيات المقترحة في (35). حيث يتم تدوير كل إحداثيات المفصل في كل إطار بشكل عشوائي حتى زاوية 13 درجة. ثم يتم تحويل إحداثيات المفصل

هذه إلى مستوى جديد مما يعطي الفيديو مظهرًا مائلًا. ويتم بعد ذلك تدوير المعلم (landmark) نسبياً مقابل المعلم الحالي أثناء مروره عبر النقاط الرئيسية لكلا اليدين. وبعد ذلك تتم إزالة السمات المكانية غير ذات الصلة إلى حد كبير عن طريق تطبيع نسب جسم المؤشر ومسافة الكاميرا وموقع الإطار مما ينتج متجه (vector) من أوضاع الجسم الطبيعية كمدخل للنموذج. ويتكون متجه (vector) وضعية كل إطار من 54 موقعًا مفصلياً يتم ترميزها بعد ذلك بمعلومات مواقعها. ويتم استخدام الترميز المكتسب بأبعاد 108 ويضاف حسب عناصره إلى متجه (vector) الوضعية. ثم يتم تغذية تسلسل الإدخال في طبقات ترميز المحول ويمر عبر وحدة الانتباه الذاتي وشبكة تغذية أمامية من طبقتين. وتتكون وحدة الانتباه الذاتي من تسعة رؤوس وست طبقات ترميز.

3.2 انتشار التضمين

يعد انتشار التضمين تقنية لرسم خريطة للسمات في مجموعة من السمات المتداخلة تسمى التضمينات. وقد استخدمنا في هذا العمل تقنية انتشار التضمين المقترحة في (34). وتنقل هذه التقنية سمات الإدخال المستخلصة باستخدام مُشفِّر المحول إلى البيانات العرضية. ثم تنتج مجموعة من التضمينات $\tilde{z}_i$ في خطوتين. أولاً يتم حساب المسافة لكل زوج من السمات (i, j)

على أنها ($d^2 = z_i - z_j^2$) ومصفوفة التجاور على أنها ($A_{ij} = exp(-d^2/\sigma^2)$ حيث σ2 هو عامل للقياس و0 = Aii لجميع i. وبعد ذلك يتم حساب لابلاس لمصفوفة التجاور على النحو التالي:

$$L = D^{-1/2} * AD^{-1/2}, D_{ii} = \sum_j A_{ij} \quad (1)$$

ومن ثم يتم الحصول على التضمينات المحسَّنة على النحو التالي،

$$P = (I - \alpha L)^{-1} \quad (2)$$

حيث I هي مصفوفة تحديد و$\alpha \in R$ هو عامل للقياس ويتم حساب التضمينات النهائية على النحو التالي،

$$\widetilde{z_i} = \sum_j P_{ij} Z_j \quad (3)$$

يؤدي انتشار التضمين إلى إزالة الضوضاء غير المرغوب فيها من متجهات السمات نظرًا لأن $\widetilde{Z_i}$ أصبحت الآن مجموعًا مرجحًا لجيرانها.

وبهدف إجراء صقل متعدد على التضمين الناتج فقد قمنا بتقييم تقنيات انتشار التسميات والشبكة النموذجية (36). ويتم إجراء عمليات تحسين النموذج وتصنيفه على مخرجات تقنية الصقل.

4. العمل التجريبي

مجموعة البيانات: لقد استخدمنا مجموعة بيانات لغة الإشارة الأمريكية على مستوى الكلمات (WLASL) لتدريب وتقييم نهجنا المقترح (37). وهي مجموعة بيانات للغة الإشارة الأمريكية تتألف من 100 إيماءة إشارة مميزة يؤدي كل منها العديد من المؤشرين مع قيام أكثر من ثلاثة مؤشرين بتنفيذ كل إشارة. وتتضمن مجموعة البيانات معلومات عن الوضعية المتخذة لأداء جميع الإشارات. وقد قسمنا البيانات في عملنا إلى ثلاث مجموعات: مجموعة أساسية تحتوي على 90 إيماءة ومجموعة تحقق تحتوي على 5 إيماءات ومجموعة فئة جديدة تحتوي على 5 إيماءات. وتم استخدام المجموعة الأساسية ومجموعة التحقق أثناء مرحلة ما قبل التدريب بينما تم استخدام المجموعة الجديدة أثناء مرحلة الاستدلال. وقمنا أثناء الاستدلال بتقسيم المجموعة الجديدة إلى مجموعات دعم واستعلام.

إعداد التجارب: يتم تحسين النماذج باستخدام مُحسِّن (SGD) أثناء مرحلة التدريب بمعدل تعلم 0.0001 وقد تم اختياره تجريبيًا. وفي كل مرة يصل فيها النموذج إلى مرحلة ثبات وهو ما يحدث عندما لا ينخفض فاقد أو خسارة التحقق لمدة 10 عمليات مرور كاملة للبيانات نقوم بتقليل معدل التعلم بعامل 10.

الجدول 1: دقة التعرف للنظام المقترح مع عدد مختلف من العينات في مجموعة العينات الداعمة. أعلى دقة مكتوبة بخط غامق وثاني أعلى درجة مكتوبة بخط مسطر.

Support set size	Without Embedding Propagation		With Embedding Propagation	
	Label Propagation	Prototypical Networks	Label Propagation	Prototypical Networks
1	72.2	67.2	70.8	68.6
5	72.4	73.4	**76.6**	72.2
10	69.8	65.4	68.8	<u>76.0</u>

النتائج والمناقشة: قمنا بتقييم النموذج المقترح باستخدام تكوينات مختلفة من خلال تغيير عدد العينات في مجموعة العينات الداعمةوتوضح النتائج في الجدول 1 تأثير انتشار التضمين على أداء النموذج في التعرف على لغة الإشارة مع توفر عينات محدودة. كما قمنا بتقييم مكونات النظام مع انتشار التضمين وبدونه لتسليط الضوء على فعالية هذه الطريقة. وكما هو موضح في الجدول فقد تم تحقيق دقة بنسبة 76.6٪ باستخدام طريقة انتشار التسميات جنبًا إلى جنب مع طريقة انتشار التضمين مقارنة بنفس الإعدادات بدون انتشار التضمين. وتم الحصول على ثاني أعلى معدل دقة 76.0٪ باستخدام الشبكات النموذجية مع انتشار التضمين مما يمثل تحسنًا بنحو 11٪ على نفس الإعدادات بدون استخدام طريقة انتشار التضمين.

ومن الواضح أيضًا أن كل من تقنيات الصقل وانتشار التسميات والشبكات النموذجية قد أدت دورها

بشكل فعال مع نموذج المحول باستخدام عدد صغير من العينات في مجموعة العينات الداعمةوعلى الرغم من أن زيادة عدد العينات قد أدى بشكل عام إلى تحسين أداء جميع التقنيات إلا أن بعض النماذج واجهت مشكلة عدم الكفاءة في التعامل مع البيانات الجديدة وهو ما قد يفسر انخفاض الأداء عند استخدام 10 عينات في مجموعة العينات الداعمة .

5. الخاتمة

اقترحنا في هذه الورقة طريقة للتعلم قليل الأمثلة للتعرف على لغة الإشارة وهي طريقة مصممة للتعميم بشكل فعال على الفئات غير المرئية سابقاً. وتقوم طريقتنا بربط السمات في مساحة الإدخال بمساحة التضمين باستخدام انتشار التضمين جنبًا إلى جنب مع تقنيات انتشار التسميات. ويتم في البداية استخراج سمات إيماءة الإشارة من إطارات الإدخال باستخدام مشفر المحول. ثم يتم تعيين هذه السمات على مساحة التضمين من خلال طريقة انتشار التضمين متبوعة بانتشار التسمية لصقل هذه التضمينات. لقد قمنا بتقييم الطريقة المقترحة باستخدام مجموعة بيانات (WLASL-100) حيث توضح النتائج التجريبية تفوق الجمع بين انتشار التضمين وانتشار التسميات مقارنة بالشبكة النموذجية. وبالنسبة للعمل المستقبلي فنحن نخطط لتقييم طريقتنا وفق مجموعات بيانات لغة الإشارة المختلفة لتقييم قدرتها على التعميم على الفئات غير المرئية سابقاً بشكل أكبر.

شكر وتقدير

يود المؤلفون أن يعربوا عن تقديرهم للدعم الذي تلقوه من الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA) وجامعة الملك فهد للبترول والمعادن (KFUPM) في إطار منحة مركز أبحاث الذكاء الاصطناعي المشترك بين (SDAIA) وجامعة الملك فهد للبترول والمعادن رقم JRC-AI-RFP-14.

المراجع

[1] E.-S. M. El-Alfy, H. Luqman, A comprehensive survey and taxonomy of sign lan- guage research, Engineering Applications of Artificial Intelligence 114 (2022) 105198.

[2] S. Alyami, H. Luqman, M. Hammoudeh, Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects, Information Processing & Management 61 (5) (2024) 103774.

[3] Y. C. Bilge, R. G. Cinbis, N. Ikizler-Cinbis, Towards zero-shot sign language recog- nition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–1doi:10.1109/TPAMI.2022.3143074.

[4] Y. Wu, T. S. Huang, Vision-based gesture recognition: A review, in: International gesture workshop, Springer, 1999, pp. 103–115.

[5] A. a. I. Sidig, H. Luqman, S. A. Mahmoud, Arabic sign language recognition using optical flow-based features and hmm, in: Recent Trends in Information and Com- munication Technology: Proceedings of the 2nd International Conference of Reli- able Information and Communication Technology (IRICT 2017), Springer, 2018, pp. 297–305.

[6] C. Neidle, A. Thangali, S. Sclaroff, Challenges in development of the american sign language lexicon video dataset (asllvd) corpus, in: 5th workshop on the represen- tation and processing of sign languages: interactions between corpus and Lexicon, LREC, Citeseer, 2012.

[7] C. Lucas, R. Bayley, Variation in sign languages: Recent research on asl and beyond, Language and Linguistics Compass 5 (9) (2011) 677–690.

[8] C. Valli, C. Lucas, Linguistics of American sign language: An introduction, Gal- laudet University Press, 2000.

[9] R. Rastgoo, K. Kiani, S. Escalera, Sign language recognition: A deep survey, Expert Systems with Applications 164 (2021) 113794.

[10] N. Cihan Camgoz, S. Hadfield, O. Koller, R. Bowden, Subunets: End-to-end hand shape and continuous sign language recognition, in: Proceedings of the IEEE inter- national conference on computer vision, 2017, pp. 3056–3065.

[11] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7784–7793.

[12] S. Stoll, N. C. Camgöz, S. Hadfield, R. Bowden, Sign language production using neural machine translation and generative adversarial networks, in: Proceedings of the 29th British Machine Vision Conference (BMVC 2018), British Machine Vision Association, 2018.

[13] M. P. Lewis, F. Gary, Simons, and charles d. fennig (eds.). 2013. ethnologue: Lan- guages of the world (2015).

[14] F. Wang, C. Li, Z. Zeng, K. Xu, S. Cheng, Y. Liu, S. Sun, Cornerstone network with feature extractor: a metric-based few-shot model for chinese natural sign language, Applied Intelligence 51 (10) (2021) 7139–7150.

[15] G. Shovkoplias, M. Tkachenko, A. Asadulaev, O. Alekseeva, N. Dobrenko,
D. Kazantsev, A. Vatian, A. Shalyto, N. Gusarova, Support for communication with deaf and dumb patients via few-shot machine learning, in: Proceedings 14th International Conference on ICT, Society and Human Beings (ICT 2021), the 18th International Conference Web Based Communities and Social Media (WBC 2021), 2021.

[16] S. Ferreira, E. Costa, M. Dahia, J. Rocha, A transformer-based contrastive learning approach for few-shot sign language recognition, arXiv preprint arXiv:2204.02803 (2022).

[17] S. Ravi, M. Suman, P. Kishore, K. Kumar, A. Kumar, et al., Multi modal spatio temporal co-trained cnns with single modal testing on rgb–d based sign language gesture recognition, Journal of Computer Languages 52 (2019) 88–102.

[18] K. M. Lim, A. W. C. Tan, C. P. Lee, S. C. Tan, Isolated sign language recogni- tion using convolutional neural network hand modelling and hand energy image, Multimedia Tools and Applications 78 (14) (2019) 19917–19944.

[19] A. Wadhawan, P. Kumar, Sign language recognition systems: A decade systematic literature review, Archives of Computational Methods in Engineering 28 (3) (2021) 785–813.

[20] S. Aly, W. Aly, Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition,

IEEE Access 8 (2020) 83199– 83212.

[21] H. Luqman, E.-S. M. El-Alfy, Towards hybrid multimodal manual and non-manual arabic sign language recognition: Marsl database and pilot study, Electronics 10 (14) (2021) 1739.

[22] P. Kumar, P. P. Roy, D. P. Dogra, Independent bayesian classifier combination based sign language recognition using facial expression, Information Sciences 428 (2018) 30–48.

[23] A. Sabyrov, M. Mukushev, V. Kimmelman, Towards real-time sign language inter- preting robot: Evaluation of non-manual components on recognition accuracy., in: CVPR Workshops, 2019.

[24] N. C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10023– 10033.

[25] B. Saunders, N. C. Camgoz, R. Bowden, Progressive transformers for end-to-end sign language production, in: European Conference on Computer Vision, Springer, 2020, pp. 687–705.

[26] W. Tao, M. C. Leu, Z. Yin, American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion, Engineering Applications of Artificial Intelligence 76 (2018) 202–213.

[27] H. Luqman, E.-S. M. El-Alfy, G. M. BinMakhashen, Joint space representation and recognition of sign language fingerspelling using gabor filter and convolutional neural network, Multimedia Tools and Applications 80 (7) (2021) 10213–10234.

[28] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3413–3423.

[29] J. Huang, W. Zhou, H. Li, W. Li, Sign language recognition using 3d convolutional neural networks, in: 2015 IEEE international conference on multimedia and expo (ICME), IEEE, 2015, pp. 1–6.

[30] O. M. Sincan, H. Y. Keles, Using motion history images with 3d convolutional networks in isolated sign language recognition, IEEE Access 10 (2022) 18608–18618.

[31] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, A. Zisserman, Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues, in: European conference on computer vision, Springer, 2020, pp. 35–53.

[32] L. Momeni, G. Varol, S. Albanie, T. Afouras, A. Zisserman, Watch, read and lookup: learning to spot signs from multiple supervisors, in: Proceedings of the Asian Con- ference on Computer Vision, 2020.

[33] R. Rastgoo, K. Kiani, S. Escalera, Zs-slr: Zero-shot sign language recognition from rgb-d videos (2021).

doi:10.48550/ARXIV.2108.10059.
URL https://arxiv.org/abs/2108.10059

[34] P. Rodríguez, I. Laradji, A. Drouin, A. Lacoste, Embedding propagation: Smoother manifold for few-shot classification, in: European Conference on Computer Vision, Springer, 2020, pp. 121–138.

[35] M. Boháček, M. Hrúz, Sign pose-based transformer for word-level sign language recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 182–191.

[36] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Ad- vances in neural information processing systems 30 (2017).

[37] D. Li, C. Rodriguez, X. Yu, H. Li, Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1459– 1469.