A Deep Learning Model To Recognize Human Movement By Using Auto Coders

Dr. Jafar Mohsen Alkheir*
Sami Haytham Abobala**

(Received 5 / 10 / 2024. Accepted 23 / 12 / 2024)

□ Abstract □

In the recent years, there has been great interest in identifying human movement using deep learning and skeletal data for its effective performance compared to models that depend on images or depth data. In this research, we presented a model that solves the problem of the size and inconsistency of data that used on deep learning model and enables these results to works in environments poor with hardware. In this paper, we proposed to use a model consisting of two stages: the first stage of the feature extraction process using autoencoder, and we made use of reducing the size of the features extracted from the training groups in order to reduce the size and complexity of the classification model represented by the second stage. The proposed model provided good results in comparison with the size and complexity of the classification model.

Keywords: deep learning, auto encoder, human action recognition, time series, LSTM.

* **Professor, Faculty of Informatics Engineering, Tishreen University, Latakia, Syria, Email: Alkheir.j@gmail.com**

****Postgraduate Student (Master) - Computer Engineering, Department of Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Latakia, Syria, Email: samiabobala@gmail.com**

استخدام المرمزات الآلية المبنية على التعلم العميق في التعرف على الحركة البشرية

د. جعفر محسن الخير*

سامي هيثم أبوبالا**

(تاريخ الإيداع 5 / 10 / 2024. قُبِل للنشر في 23/ 12 / 2024)

□ ملخّص □

في السنوات الأخيرة ازداد الاهتمام في مجال التعرف على الحركة البشرية باستخدام التعلم العميق وبيانات الهيكل العظمي لما حققته من فعالية وأداء عالي مقارنة بالنماذج التي تعتمد على الصور أو بيانات العمق. في هذا البحث قدمنا نموذج يحل مشكلة حجم وعدم تناسق البيانات المستخدم ضمن نماذج التعلم العميق والتي تسمح لتلك النماذج بالتدريب حتى في بيئات فقيرة بالعتاد الصلب. اقترحنا في هذا البحث استخدام نموذج مؤلف من مرحلتين الأولى عملية استخلاص السمات باستخدام المرمزات الآلية Auto Encoder واستفدنا من تقليل حجم السمات المستخلصة من مجموعات التدريب من أجل تقليل حجم وتعقيد نموذج التصنيف والذي تمثله المرحلة الثانية. قدم النموذج المقترح نتائج جيدة بالمقارنة مع حجم وتعقيد نموذج التصنيف. ويمكن القول ان هذا البحث يفتح الباب أمام المزيد من الأبحاث في مجال استخدام المرمزات الآلية للمساعدة في تصنيف السلاسل الزمنية المتعلقة بالحركة البشرية وإدخال تطويرات على بنى المرمزات الآلية والمصنف المستخدم.

الكلمات المفتاحية: التعلم العميق، التعرف على الحركة البشرية، المرمزات الآلية، معالجة السلاسل الزمنية، LSTM.

* أستاذ -كلية الهندسة المعلوماتية- جامعة تشرين- اللاذقية – سورية . Email j.Alkheir@gmail.com

** طالب دراسات عليا (ماجستير) – قسم هندسة الحاسبات والتحكم الآلي – كلية الهندسة الميكانيكية و الكهربائية – جامعة تشرين – اللاذقية – سورية Email: samiabobala@gmail.com

مقدمة:

في السنوات الأخيرة جرى استخدام التعرف على الحركة البشرية بشكل واسع النطاق في مجالات مختلفة، مثل فهم السلوك البشري، كاميرات المراقبة، التحكم في السيارات الحديثة وأي شكل من أشكال التواصل بين الحاسب والإنسان. مع تطور هذه الاستخدامات وتعقيدها دعت الحاجة إلى تطوير الخوارزميات المستخدمة ضمن هذا المجال، حيث قدم التعلم العميق أداءً ممتازاً في تحسين دقة التعرف على الحركة بالمقارنة مع الطرق التقليدية. من بين أنظمة أو طرق التعلم العميق كان أداء النظم المعتمدة على بيانات الهيكل العظمي أكثر دقة و فعالية في حل هذه المشكلة. هذا التحسن كان على حساب سرعة التنفيذ من جهة وضخامة حجم النموذج المقترح من جهة أخرى [1]،[2] ،[3] ،[4] ، [5] مما جعل النماذج المقترحة واستخدام التعلم العميق ككل غير مناسب لجميع السناريوهات –النظم المضمنة كمثال– . ففي التطبيقات العملية بعيداً عن بيئة المخابر المتفوقة على صعيد العتاد يجب أن يعمل النظام أو النموذج المقترح بشكل فعال على أقل عتاد صلب متوفر بغض النظر إذا كانت الحركة المراد التعرف عليها مركبة أو يقوم أكثر من شخص بإجرائها بزوايا مختلفة للكاميرة قدر المستطاع. لحل المشاكل السابقة وأهمها حجم النموذج وأدائه في مختلف الظروف. قمنا باقتراح نظام مؤلف من مرحلتين 1. عملية فلترة السمات من الضجيج والقراءات الزائدة 2. مرحلة التعرف على السمات وعلى الحركة.

على وجه التحديد قمنا بدراسة مجموعات البيانات المستخدمة في هذا المجال بشكل مفصل وأخذنا بعين الاعتبار عدد الحركات الموجودة في مجموعة البيانات و نوعها و عدد المؤدين و أيضاً شروط تأدية الحركة من نوع الكاميرا المستخدمة و طريقة تنصيبها. خلال التعامل مع البيانات في هذه المرحلة واجهتنا مشاكل تتعلق ببيئة التجربة من حيث ضعف العتاد المادي، ومن ثم تم إجراء عمليات تحويل على إحداثيات الهياكل العظمية في مجموعة البيانات من أجل تقليل التغيرات العشوائية الناتجة عن العامل البشري أثناء تأدية الحركة. ليتم بعدها التخلص من الضجيج الموجود في مجموعة البيانات باستخدام المرمزات الآلية Auto Encoders، حيث تم الاستفادة من فقدان السمات الغير أساسية والتي تعتبر عادة من مساوئ هذه التقنية. لقد تم توحيد أطوال سلاسل الحركة لكل مقطع من أجل تقليل التشويش في دخل الشبكة العصبونية ليتم بعدها إدخال نتيجة هذه المرحلة على شبكة الذاكرة طويلة و قصيرة الأمد LSTM للقيام بعملية التعرف. تجدر الإشارة أنه أثناء مرحلة الترميز الآلي تم استخلاص نقطة الاختناق من المرمز وإضافتها كدخل إلى مرحلة التعرف مما يعزز عملية المفاضلة بين الحركات المختلفة.

مشكلة البحث:

يواجه معظم الباحثون في مجال التعرف على الحركة البشرية عدة مشاكل أهمها تعقيد النموذج المقترح مما يؤدي إلى زيادة حجمه وزمن تدريبه والموارد المستخدمة في تشغيله. من أهم المشاكل التي تواجه الباحثين هي ارتباط النموذج المدرب بشكل وثيق بقاعدة البيانات المستخدمة في التدريب لأن لكل قاعدة بيانات طريقة مختلفة في تسجيل الحركة وزمن مختلف للحركة الواحدة. بالإضافة إلى ذلك تواجه هذا النوع من الأبحاث مشكلة هامة وهي الضجيج الناتج عن القراءات الزائدة للقيم داخل مجموعة البيانات.

أهمية البحث وأهدافه:

يهدف هذا البحث إلى بناء قاعدة بيانات عامة يمكن استخدامها بغض النظر عن النموذج المستخدم الأمر الذي يؤدي إلى تقليل تعقيد النماذج وحجمها وزمن تدريبها والموارد المستخدمة في تشغيلها.

تجدر الإشارة إلى أن أغلب الأبحاث تحاول حل المشاكل السابقة قبل مرحلة التدريب، أما في هذا البحث فقد جزئنا حل المشكلة لمرحلتين الأولى قبل التدريب والثانية أثناء تصميم نموذج التعلم العميق. في هذا البحث أجرينا معالجة أولية للبيانات بحيث قللنا حجم النموذج النهائي وعالجنا مشكلة القراءات الزائدة.

طرائق البحث ومواده:

أعتمد هذا البحث على المنهج التجريبي مع تجزيئ مشاكل البحث ورسم خطوط عريضة مسبقة للتقنيات التي من الممكن أن تقدم حلاً لها. تم استخدام بيئة الحوسبة السحابية CoLab المقدمة من شركة Google وبالإضافة إلى استخدم بيئة الـ TensorFlow حيث تم التعامل معها باستخدام مكتبة Keras. تم اختيار مجموعة بيانات NTU RGBD [43] حيث تعتبر أشمل قاعدة بيانات في هذا المجال وإخذنا منها معلومات الهيكل العظمي فقط.

الدراسات المرجعية:

يمكن تقسيم الدراسات المرجعية إلى مجالين وهما: الدراسات المنجزة في مجال معالجة الصورة والدراسات المنجزة في مجال بناء مجموعة البيانات التي تستخدم في عملية تدريب النماذج واختبارها والتي تعتبر الخطوة الأولى والأهم في تطبيقات أنظمة معالجة الصور.

في مجال الصور:

خلال السنوات الأخيرة، تم التركيز بشكل كبير على تطبيقات التعلم العميق في مختلف المجالات وذلك للتفوق الذي حققته مقارنة بالخوارزميات التقليدية، حيث جرى استخدامها بشكل واسع على وجه الخصوص في مجال معالجة الصورة. لذلك تعتبر البيانات المستخلصة من الصور أو الفيديوهات باستخدام الشبكات العصبية الالتفافية CNN حجر الأساس في التعرف على الحركة البشرية المستخلصة من البيانات البشرية.

اعتماداً على نوع البيانات الخام، يمكن تقسيم الأبحاث في مجال التعرف على الحركة البشرية إلى أبحاث معتمدة على الصور الرقمية RGB [6][7][8] وأبحاث اعتمدت على دمج الصور الرقمية مع العمق RGBD والذي يعتبر الأكثر انتشاراً من النوع السابق [9] [10] [11] [12] [13]

تؤمن الصور التي تحتوي خاصية العمق RGBD عزل فعال بين الغرض من جهة والمتغيرات في البيئة وخلفية الصورة. لذلك تكون عملية تجزئة الأشياء باستخدام هذه الصور سريعة جدا وفعالة. وبناءً عليه قدمت هذه التقنية إمكانية بناء أنظمة فعّالة للتعرف على الحركة البشرية في الزمن الحقيقي والتي أعطت دقة عالية في التعرف على الحركة البشرية خلال زمن معالجة قليل وتعقيد أقل في البرامج. بالإضافة إلى ذلك، تعد من الطرق المنتشرة بشكل كبير في مجال التعرف على الحركة البشرية. [14] [15] [16]

لاحقاً تم إجراء أبحاث واسعة في مجال التعرف على الحركة البشرية باستخدام البيانات المعتمدة على حركة الهيكل العظمي. يمكن تصنيف هذه الأبحاث اعتماداً على الطريقة التي تم استخلاص بيانات الهيكل العظمي إلى قسمين أساسيين: الأول بيانات مستخلصة يدوياً والثانية بيانات مستخلصة باستخدام التعلم العميق. كمثال على النوع الأول

هناك أبحاث قامت بوصف الهيكل العظمي ثلاثي الأبعاد باستخدام Lie group [17] أو vector−valued function [18]. لكن أداء هذه الطرق كان محدوداً للغاية. في الوقت الحالي، تعتبر الطرق التي تعتمد على التعلم العميق أكثر انتشاراً، حيث حققت نتائج ممتازة لقدرتها على استخلاص مجموعة أكبر من السمات وبشكل تلقائي. يمكن تقسيم هذه الطرق إلى ثلاث أصناف هي:

طرق معتمدة على شبكات الـ RNN [23] [22] [21] [20] [19]

وطرق معتمدة على شبكات الـ CNN [27] [26] [25] [24]

وطرق معتمدة على شبكات الـ GCN [28] [29] [30] [31] [32] [33] [34]

تعتمد بنية الشبكات العصبية التكرارية RNN على إدخال حالة الغرض السابقة _أو إحداثياته في حالتنا_ كدخل للحالة الآنية وهكذا مما يؤمن معالجة أكثر فعالية في المجالات التي تطلب أخذ الزمن بعين الاعتبار وأهمها التعرف على الحركة البشرية. تم تطوير هذا النوع من الخوارزميات إلى LSTM و GRU ليتمكن من حل المشاكل التقنية التي واجهته والمتمثلة في تلاشي الانحدار وإدخال بيانات من مراحل سابقة غير ذات صلة والتي أدت إلى انخفاض مستوى دقة خوارزمية RNN التقليدية في مجالات الترجمة والتعرف على النصوص والتعرف على الحركة البشرية. لكن مع ذلك بقيت مشكلة إدخال المعلومات المكانية في مجالات التعرف على الحركة البشرية .[36] [35]

مؤخراً أقترح الباحثون خوارزمية تعتمد على شبكتي RNN من أجل معالج المعلومات المكانية والمعلومات الزمنية في عملية التعرف على الحركة البشرية [37] Hong and Liang. بالمقابل أجرى باحثون أخرون معالجة مبدئية على بيانات الهيكل العظمي المكانية مثل [38] Jun and Amir الذين قاموا بتطبيق خوارزمية مسح جديدة من أجل إعادة ترتيب مواقع العقد في محاولة للحصول على روابط مكانية وزمنية أكثر فائدة بالإضافة إلى إزالة التشويش الناتج عن ترتيب بيانات كاميرا Kinect للعقد بدون الأخذ بعين الاعتبار العلاقة التبعية الحركية بين المفاصل المتجاورة. باستخدام شبكات الـ LSTM و البنية الشجرية الجديدة كدخل يؤمن تمثيل العلاقات المكانية بشكل جيد. يتم تحديث بيانات الشبكة إذا كانت الشجرة في الخطوة الجديدة تمثل ترابط مكاني معقول بالنسبة للدخل السابق وهذا مستوحى من الشبكات الالتفافية و التي تعتبر مناسبة جداً لنمذجة العلاقات المكانية. لاحقاً قام كل من [39] Chunyu and Baochang باستخدام الشبكات الالتفافية وشبكة RNN مع ميزة التثقيل من أجل الحصول على نمذجة أفضل للترابط الزماني المكاني المعقد بين عقد الهيكل العظمي. التثقيل المكاني يستخدم من أجل تقدير الإطارات المهمة للحركة خلال الزمن، ليتم بعدها استخدام شبكة الـ CNN من أجل تحليل ونمذجة الترابط المكاني بين العقد المختلفة الموجودة ضمن الإطارات المرشحة من شبكة RNN. في أبحاث أخرى [40] تم استخدام شبكة RNN مثقلة من أجل الحصول على تمثيل مكاني أفضل للسمات المستخلصة من الهيكل العظمي ومن ثم استخدام شبكة LSTM متعددة الطبقات من أجل الحصول على ترابط زماني.

الشبكات الالتفافية البيانية GNC:

تقوم الشبكات الالتفافية البيانية بتطوير مفهوم الشبكات الالتفافية لتستطيع التعامل مع بنية البيان، وهنالك طريقتين لبناء هذا النوع من الشبكات هما التمثيل المكاني والتمثيل الطيفي. في التمثيل المكاني يتم تطبيق الفلاتر الالتفافية على عقد البيان وعلى العقد المجاورة. على عكس التمثيل المكاني، يعتبر التمثيل الطيفي للبيان كشكل من أشكال التحليل الطيفي حيث يعتمد على تحويلات لابلاس لمليء متجهات البيان.

مجموعة البيانات المستخدمة في هذا المجال:

مجموعة بيانات UTKinect–Action3D:

هذه المجموعة مكونة من عشر حركات بشرية في بيئة معدة داخل المنزل، تم أخذ تسلسل لقراءات عقد الهيكل العظمي باستخدام كاميرا Kinect وحيدة. الكاميرا تملك مدى فعلي بين المتر والثلاثة أمتار ونصف. الصور الملتقطة وبيانات العمق تم التقاطها بسرعة 30 إطار بالثانية FPS. دقة صور العمق الملتقطة هي 240 x 320بينما دقة صور RGB 480x640 [41].

أنواع الحركات العشرة تتضمن: المشي والجلوس والنهوض والتقاط شي من الأرض وحمل الأشياء والدفع والشد والتلويح باليد والتصفيق. كل حركة تم تمثيلها من قبل عشرة أشخاص مرتين: تسعة رجال وامرأة. أحد هؤلاء الأشخاص إيسر. بناءً على ما سبق يوجد في قاعدة البيانات هذه 6220 إطار تكون 200 حركة. طول الحركة الواحدة يتراوح بين 5 إطارات و120 إطار.

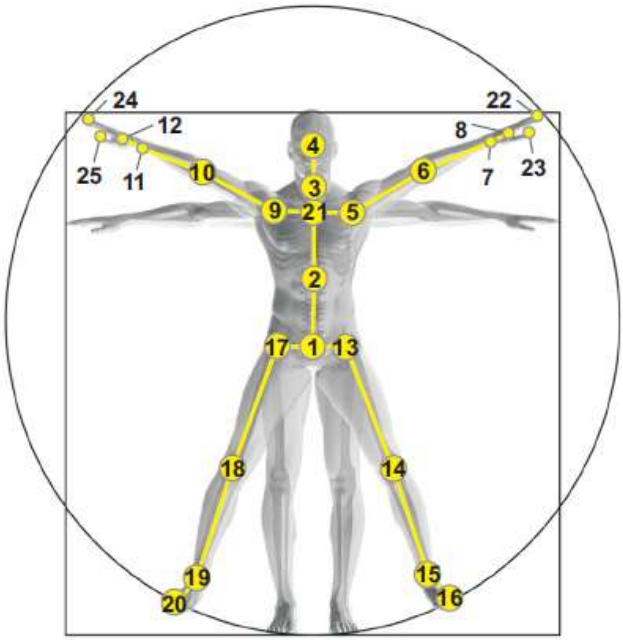
D2. SYSU 3D Human–Object Interaction Dataset

تم بناء قاعدة البيانات هذه من قبل جامعة Yat–sen عام 2015. ركزت قاعدة البيانات هذه على التفاعل بين البشر والأشياء فقد تضمنت 40 شخصاً يقومون بإثني عشر حركة مختلفة منها شرب الماء والتململ والتحدث على الهاتف واللعب بالهاتف وارتداء حقيبة ظهر وتوضيب حقيبة الظهر والجلوس على الكرسي وتحريك كرسي وإخراج المحفظة من الجيب والتأشير باليد والتلويح باليد. لكل حركة، يقوم شخص بالتفاعل مع أحد الأغراض الستة التالية: هاتف، حقيبة، كرسي، محفظة، ممسحة ومكنسة، لذلك تحتوي قاعدة البيانات على 480 فيديو بالمجمل. تم استخدام كاميرا Kinect من أجل التقاط صور RGB وبيانات عقد الهيكل العظمي. يدعي منشئ قاعدة البيانات هذه أنها تحتوي على تحديات جديدة لا تملكها قواعد بينات أخرى حيث أن بعض الحركات مشابهة لحركات التفاعل بين الأشخاص والأغراض وعدد الأشخاص المشاركين في أداء الحركات أكبر بكثير من بقية قواعد البيانات الأخرى[42] .

NTU RGB+D dataset.وهي مجموعة البيانات المستخدمة في هذا البحث كونها أكثر شمولاً من حيث عدد المؤدين و عدد الحركات و الكاميرات المستخدمة في عملية التقاط الحركة

تم استخدام كاميرا Kinect من النسخة الثانية في بناء قاعدة البيانات هذه. وتم التقاط بيانات مختلفة باستخدام هذه الكاميرا تتضمن: صور عمق وتمثيل ثلاثي الأبعاد لبيانات مفاصل الهيكل العظمي وصور RGB عادية وصور أشعة تحت الحمراء. للحفاظ على كل المعلومات تم تطبيق خوارزمية ضغط لا تؤدي إلى فقدان في البيانات على كل إطار. دقة صور كل إطار هي 512 x 424. معلومات الهيكل العظمي تتألف من إحداثيات ثلاثية الأبعاد لخمس وعشرين مفصل أساسي في جسم الإنسان والتي تساعد على تتبع حركة الجسم في المشهد، وكذلك تم إضافة معلومات البيكسلات والعمق لكل مفصل من مفاضل الجسد في الإطار الواحد كما هو موضح في (الشكل1). تم تسجيل فيديوهات الـ RGB بدقة 1080x1920. تملك قاعدة البيانات 60 حركة مختلفة تم تقسيمها إلى ثلاث مجموعات أساسية: 40 حركة مرتبطة بالنشاطات اليومية (الأكل، الشرب، المطالعة...) وتسع حركات مرتبط بالصحة (السعال، التعب..) وإحدى عشر حركة اعتيادية (الركل، الضرب، العناق...). الأشخاص المشاركون في إنشاء قاعدة البيانات عددهم أربعين مشارك كانت أعمارهم تتراوح بين العشر سنين والخمس والثلاثون سنة. تم استخدام ثلاث كاميرات في تسجيل مقاطع الحركات ثبتت على زوايا −45 و0 و +45 درجة. طلب من كل شخص أن يقوم بالحركة مرتين، الأولى في مواجهة الكاميرا المثبتة على زاوية +45 بينما الأخرى مقابل الكاميرا −45 وبهذه الطريقة تم التقاط مشهدين

أماميين لكل حركة يميني ويساري [43]. تم حفظ قراءات إحداثيات مفاصل الهيكل العظمي وبيانات العمق ضمن ملفات نصية على شكل قيم عددية لكل مفصل _*25 سطر*_ لكل إطار زمني frames وهذا أدى إلى مشكلتين أساسيتين: 1. وجود بيانات غير مرغوبة مثل بيانات العمق 2. إختلاف طول الملفات النصية، حيث يختلف عدد الإطارات frames من ملف لآخر حسب المؤدي أو طوال المقطع المسجل.



الشكل 1 تمثيل الهيكل العظمي في قاعدة بيانات NTU RGB-D

1- قاعدة العمود الفقري 2- منتصف العمود الفقري 3- رقبة 4 -رؤوس 5- كتف أيسر 6- كوع أيسر 7- معصم أيسر 8- اليد اليسرى 9 الكتف الأيمن 10 الكوع الأيمن 11 -المعصم الأيمن 12ا-ليد اليمنى 13- الورك اليسرى 14-الركبة اليسرى 15- الكاحل اليسرى 16- القدم اليسرى 17- الورك الأيمن 18- الركبة اليمنى 19- الكاحل الأيمن 20- القدم اليمنى 21- العمود الفقري 22-طرف اليد اليسرى 23 - الإبهام الأيسر 24 - طرف اليد اليمنى 25-ابهام اليد اليمنى

تحضير ومعالجة البيانات:

قبل البدء في مرحلة التدريب والمعالجة الأولية للبيانات، تم تقسيم مجموعة البيانات على الشكل التالي: قسمنا الأربعين مؤدي إلى مجموعتين، مجموعة تدريب ومجموعة اختبار تتألف الواحدة منها من 20 شخص. حيث استخدم المؤدون 1، 2، 4، 5، 8، 9، 13، 14، 15، 16، 17، 18، 19، 25، 27، 28، 31، 34، 35، 38 كمجموعة تدريب والباقي كمجموعة اختبار.

من جهة أخرى ومن أجل الحصول على نتائج أكثر واقعية تم استخدام تسجيلات الكاميرا 2 و3 من أجل التدريب بينما استخدمنا تسجيلات الكاميرا 1 من أجل الاختبار. نتج عن ذلك 40320 عينة تدريب و16560عينة اختبار.

تجدر الإشارة أن التقسيم السابق يساعد على التخلص من ظاهرة overfitting في النموذج المقترح.

تم إجراء عمليات نقل وتحويل على إحداثيات المفاصل كي تكون أكثر تجانساً، حيث تم نقل مركز الإحداثيات من مركز الكاميرا إلى المفصل رقم 1 من مفاصل الهيكل العظمي (الشكل1) ومن ثم نسب باقي أحداثيات الملف

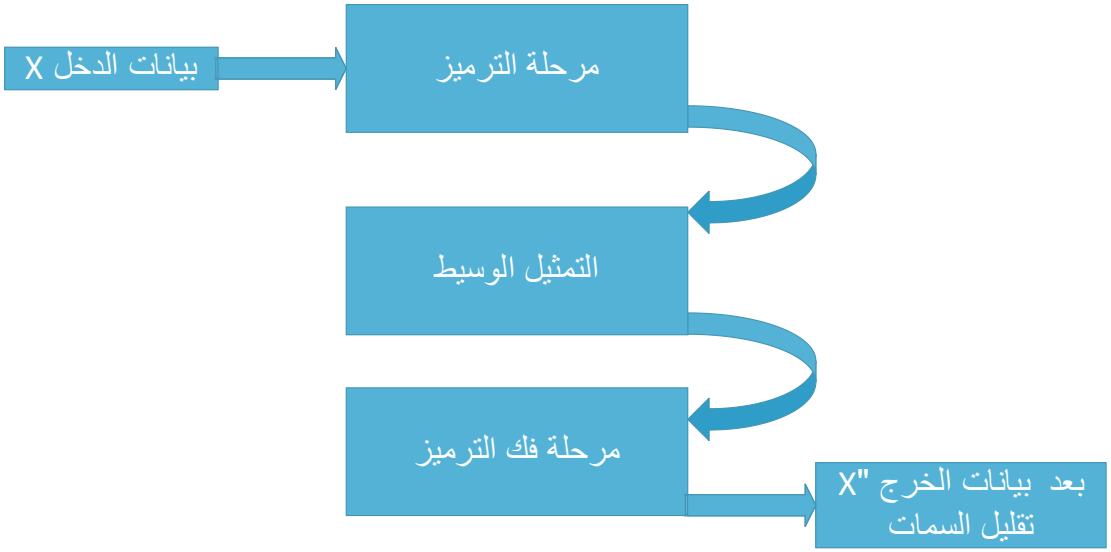
لإحداثيات المفصل الأول وهذا أدى إلى حل مشكلة التباين في مركز الإحداثيات بين تسجيلات الحركات الناتج عن اختلاف بعد الشخص عن الكاميرا.

ومن ثم تم العمل على حل مشكلتي التفاوت في أطوال مقاطع الحركة المسجلة في مجموعة البيانات وكثرت القراءات التي من الممكن أن تؤدي إلى زيادة زمن التدريب والحصول على دقة تعلم منخفضة جداً عن طريق أخذ متوسط طول الإطارات ضمن مجموعة البيانات و من ثم استخدام الZero padding لإكمال النقص الحاصل و من أجل مشكلة القراءات الزائدة ضمن مجموعة البيانات تم استخدام المرمزات الآلية لتقليل عدد القراءات و التركيز على المفيدة منها.

المرمزات الآلية:

المرمزات الآلية اقترحت في بداية الأمر عام 1980 [44] وكانت الغاية منها إيجاد تمثيل أبسط للبيانات باستخدام سمات بعدد أقل وبفعالية أكبر. لاحقا في عام 2006 [45] تم انشاء المرمزات الآلية باستخدام شبكات التعلم العميق وحققت نتائج أكثر دقة في عملية استخلاص السمات. يتألف المرمز الآلي من مرحلتين أساسيتين: 1. الأولى مرحلة الضغط أو الترميز 2. مرحلة إعادة توليد الخرج أو فك الضغط. كما هو موضح في (الشكل 2)

مرحلة الضغط والتي تتألف من شبكة تغذية أمامية مكونة من عدة طبقات يكون هدف المرمز اختزال السمات المدخلة لتصل إلى طبقة التمثيل representation layer والتي تعتبر خرج المرحلة الأولى. المرحلة الثانية فك الضغط، وهي عبارة عن شبكة عصبية تتخذ من طبقة التمثيل دخلاً لها بهدف الحصول على دخل المرحلة الأولى مجدداً. وبالتالي يمكننا القول إن خرج المرحلة الثانية هو تمثيل شبه مطابق لدخل المرحلة الأولى لكن مع اختصار أكبرعدد ممكن من السمات وهنا تأتي فائدة هذه التقنية. بكلام آخر يمكن اعتبار أعاد المرمز الآلي إنتاج دخله بعدد سمات أقل وتجدر الإشارة إلى أن عرض طبقة التمثيل يحدد كمية السمات التي يمكن حذفها أو الحفاظ عليها مع مراعات الحصول على الخرج المطلوب.

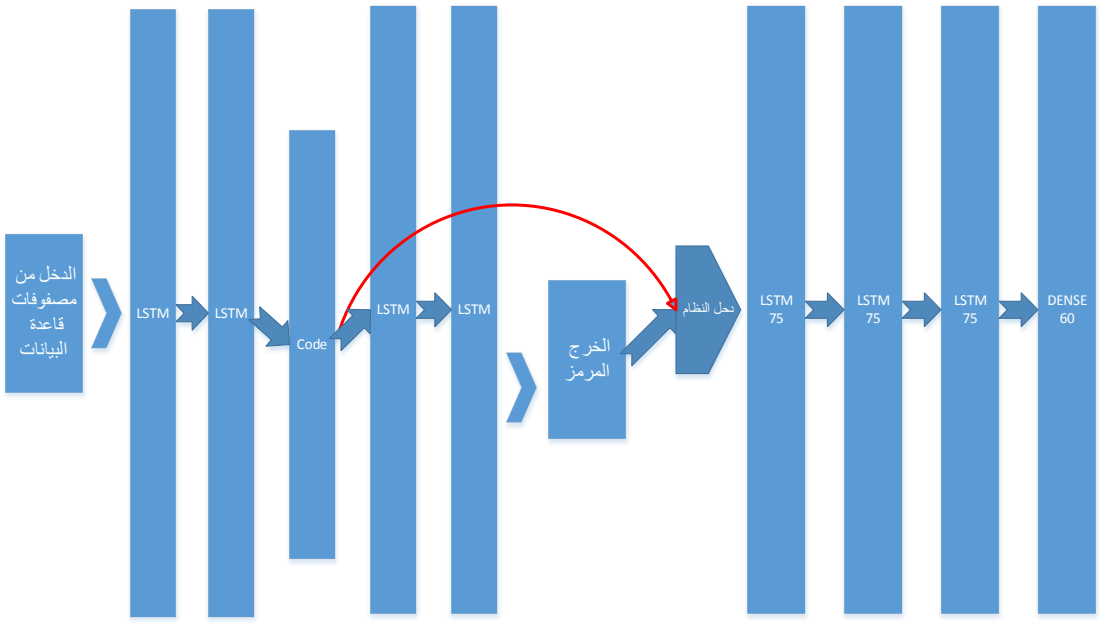


الشكل 2 مخطط عمل المرمز الآلي

النموذج المقترح

يتألف النموذج المقترح من قسمين رئيسيين الأول هو المرمز الآلي الذي يقوم بضغط الدخل وإعادة تشكيله ليعطي خرج بطول ثابت وبضجيج أقل حيث يقوم باستخلاص السمات الأساسية. يتضمن المرمز مرحلتين الأولى مرحلة الترميز وهي مؤلفة من طبقتي LSTM تستخدم لضغط سلسلة الدخل وصولاً إلى طبقة التمثيل والثانية فك الترميز وهي مؤلفة من طبقتي LSTM ايضاً تستخدم لفك الضغط انطلاقا من مرحلة الترميز وصولاً لخرج مشابه قدر الإمكان للدخل لكن بأقل عدد سمات ممكن ليتم استخدام خرج المرمز كدخل للمرجلة التالية وهي مرحلة التصنيف.

تقوم مرحلة التصنيف بأخذ خرج المرحلة السابقة بالإضافة إلى قيمة طبقة الترميز representation layer كدخل وهذه المرحلة مؤلفة من ثلاث طبقات LSTM وطبقة كاملة الاتصال DENSE حيث تقوم طبقات الـ LSTM بعملية استخلاص السمات من الدخل والطبقة الأخيرة تقوم بعملية التصنيف. (الشكل 3)



الشكل 3 مخطط النموذج المقترح

مرحلة التدريب:

تم تقسيم مرحلة التدريب إلى قسمين أساسين: أولاً تدريب المرمز الآلي فقط، حيث يقوم بأخذ العينة وضغط سماتها ويحاول التقليل قدر الإمكان من الخطأ بين الدخل والخرج. المرحلة الثانية هي تدريب المصنف وهنا نقوم بحفظ أوزان المرمز الآلي وتنحصر عملية التدريب وتحديث الأوزان بطبقات المصنف فقط.

سيناريوهات التدريب:

قمنا بإجراء ثلاث تجارب في هذا البحث: 1. تدريب النموذج بدون مرحلة المرمز الآلي أي المصنف بشكل مباشر، وهذه السيناريو مهم جداً لبيان القيمة المضافة التي قدمها المرمز الآلي في هذا المجال 2. تدريب النموذج بدون مرحلة فك الترميز أي أخذ التمثيل الوسيط بشكل مباشر. 3. تدريب النموذج كاملاً كما هو موضح في الشكل 2.

معايير تقييم الأداء:

استخدمنا في هذا البحث المعايير القياسية التالية:

1. الدقة accuracy وهي النسبة المئوية للتصنيفات الصحيحة بالنسبة لمجمل التصنيفات.
2. مصفوفة الإلتباس ويتم حسابها وفق كل حركة مراد التعرف عليها ويمكن تبسيط معادلة حساب هذه المصفوفة إلى المعادلة التالية:

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

المعادلة 1 حساب مصفوفة الإلتباس

حيث TP تعني عدد الحالات التي تم تصنيفها بشكل صحيح على أنها إيجابية. و FP عدد الحلات التي تم تصنيفها بشكل خاطئ على أنها أيجابية. أما TN عدد الحالات التي تم تصنيفها بشكل صحيح على أنها سلبية وFN عدد الحالات التي تم تصنيفها بشكل خاطئ على أنها سلبية.

3. F−Score: و هو مقياس توافقي بين الدقة و الاسترجاع يهدف إلى تحقيق توازن بين تجنب التنبؤات الكاذبة والاكتشاف الجيد للحالات الإيجابية.

النتائج والمناقشة:

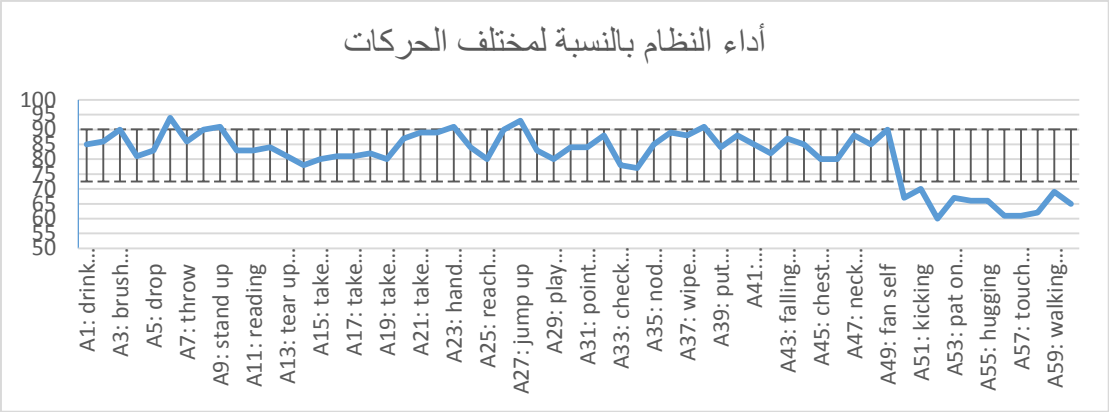
كون هدف النظام هو التعرف على مختلف الحركات البشرية المذكورة في قاعدة البيانات فقد حددنا نسبة التعرف لكل حركة على حدى و من ثم بنينا ملاحظاتنا على الشكل العالم للنتائج التي يمكن تلخيصها على الشكل التالي:

تمت التجربة الأولى بإستخدام المصنف فقط بدون مرحلة المرمز الآلي و أعطى نتائج سيئة للغاية حيث بلغت نسبة التعرف 10% فقط وهذا متوقع جداً كون المصنف بسيط مقارنة مع حجم و تنوع البيانات من جهة و تعقيد النماذج المذكورة في مختلف المراجع الموضحة في هذا المجال والتي لا تعتمد على المرمزات الآلية بل معالجة ضجيج البيانات بشكل مباشر ضمن المصنف.

أما في التجربة الثانية عند إستخدام النموذج بدون مرحلة فك الترميز، حقق النموذج نسبة تعرف قريبة من التجربة الأولى كون هذه التجربة بالذات تتطلب تعديل حجم طبقة التمثيل الوسيط و بالتالي الإستغناء فعلياً عن بعض السمات أي مفاصل الجسم و هذا يحتاج إلى دراسة السمات الأكثر أهمية ولم نتابع في بحثنا في هذا الإتجاه.

عند إستخدام النموذج الموضح في الشكل 3 كانت النتائج كمايلي.

- حقق النظام نسبة تعرف تبلغ 80.4% لمجمل الحركات.
- حقق النظام نسبة تعرف تبلغ 84.8% في الحركات التي يقوم بها مؤدي واحد فقط.
- نلاحظ من الخط البياني في الاسفل انخفاض واضح في دقة النظام عند الحركات المؤدات من قبل شخصين.
- ويمكن تلخيص النتائج كمايلي:
  - أعلى دقة مسجلة حسب الحركة كانت للإلتقاط و القفز على التوالي 94% و 93%
  - أقل دقة مسجلة حسب الحركة كانت لإعطاء شيئ و دفع شخص على التوالي 61% و 60%
  - أعلى F score مسجلة حسب الحركة كانت للإلتقاط و القفز على التوالي 95% و 94%
  - أقل F score مسجلة حسب الحركة كانت لإعطاء شيئ و دفع شخص على التوالي 66% و 62%



الشكل 4 أداء النظام بالنسبة لمختلف الحركات

ويمكن مقارنة نتائج هذا البحث مع أهم الأبحاث الحالية و السابقة في هذا المجال كما هو موضح في الجدول التالي:

الجدول 1 مقارنة أداء النظام مع الداراسات المرجعية من نفس النوع

الطريقة	الدقة
Deep LSTM [43]	60.70%
Part aware LSTM [43]	62.90%
ST-LSTM [46]	69.20%
STA-LSTM [47]	73.40%
AV-LATM [48]	79.20%
Auto-encoder LSTM النموذج المقترح في هذا البحث	**80.40%**
ST-GCN [49]	81.50%
AGC-LSTM [50]	89.20%
2S-AGCN [51]	88.50%
SHIFT-GCN [52]	90.70%
CTR-GCN [53]	92.40%

الاستنتاجات والتوصيات:

في هذا البحث قمنا باقتراح نظام تعرف على الحركة البشرية باستخدام التعلم العميق بناءً على بيانات الهيكل العظمي للإنسان واستخدمنا المرمزات الآلية كمقاربة جديدة لحل بعض المشاكل المتعلقة بهذا المجال. حيث قمنا بإنشاء نموذج مؤلف من مرحلتين الأولى هي المرمز الآلي الذي ساعدنا على ضغط سمات سلاسل الدخل من أجل تقليل الضجيج الناتج عن القراءات الكثيرة المأخوذة من كاميرات التقاط الحركة وقد أعطت هذه المرحلة تمثيل لسلاسل الحركة بعدد سمات محدد ليتم أخذ خرج هذه المرحلة كدخل لمرحلة التصنيف.

حقق البحث نسبة تعرف جيدة جداً مقارنة مع آخر الأبحاث الحالية في هذه المجال المعتمدة على أشكال جديدة من شبكات التعرف العميق (الشبكات المعتمدة على البيان على وجه التحديد) لكن كان حجم وتعقيد النموذج الناتج قليل نسبياً بالمقارنة مع نتائج هذه الأبحاث وهذا هدف البحث . يمكن في المستقبل العمل على تطوير النموذج المقترح من

خلال اختيار مرمز بيانات يأخذ بعين الاعتبار العلاقات المكانية والزمنية بين سلاسل الحركة من جهة وتطوير مرحلة التصنيف من خلال استخدام الشبكات المعتمدة على البيان وبدمج كلا التقنيتين.

References:

[1]Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset," in 10th Eurographics Workshop on 3D Object Retrieval, 2017.

[2] G. Devineau, W. Xi, F. Moutarde, and J. Yang, "Convolutional neural networks for multivariate time series classification using both interand intra-channel parallel convolutions," in Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP'2018), 2018.

[3] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatialtemporal attention res-tcn for skeleton-based dynamic hand gesture recognition," gesture, vol. 30, no. 5, p. 3, 2018.

[4] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multistream networks exploiting pose, motion, and appearance for action classification and detection," in Computer Vision (ICCV), 2017 IEEE International Conference on Computer Vision. IEEE, 2017, pp. 2923–2932.

[5] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in CVPR 2018, 2018.

[6] Dawn, D.D.; Shaikh, S.H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis. Comput. 2016, 32, 289–306.

[7] Liu, A.; Xu, N.; Nie, W.; Su, Y.; Wong, Y.; Kankanhalli, M.S. Benchmarking a Multimodal and Multiview and Interactive Dataset for Human Action Recognition. IEEE Trans. Syst. Man Cybern. 2017, 47, 1781–1794.

[8]Fernando, B.; Gavves, E.; Oramas, M.J.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.

[9] Yang, X.; Tian, Y.L. Super Normal Vector for Activity Recognition Using Depth Sequences. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 804–811.

[10] Li, M.; Leung, H.; Shum, H.P.H. Human action recognition via skeletal and depth based feature fusion. In Proceedings of the Motion in Games 2016, Burlingame, CA, USA, 10–12 October 2016; pp. 123–132.

[11] Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. J. Real-Time Image Process. 2016, 12, 155–163.

[12] Zhang, J.; Li, W.; Ogunbona, P.; Wang, P.; Tang, C. RGB-D-based action recognition datasets. Pattern Recognit. 2016, 60, 86–105.

[13] Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In Proceedings of the ECCV, Amsterdam, The Netherlands, 8–16 October 2016; pp. 816–833.

[14] Oreifej, O.; Liu, Z. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In Proceedings of the Computer Vision and Pattern Recognition, Portland, OR, USA,
23–28 June 2013; pp. 716–723.

[15] Yang, X.; Tian, Y.L. Effective 3D action recognition using EigenJoints. J. Vis. Commun. Image Represent. 2014, 25, 2–11.

[16] Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. J. Real-Time Image Process. 2016, 12, 155–163.
[17] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie Group," in CVPR 2014.
[18] B. Fernando, E. Gavves, J. Oramas M., A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in CVPR 2015.
[19]S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton da-ta," in CVPR 2016.
[20] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3d human action recognition," in ECCV 2016.
[21] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adap-tive recurrent neural networks for high performance human action recognition from skeleton data," in ICCV 2017.
[22] M. Rhif, H. Wannous, and I. R. Farah, "Action recognition from 3d skeleton sequences using deep networks on lie group features," in ICPR
[23] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang, "Relational net-work for skeleton-based action recognition," in ICME 2019.
[24] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in ICMEW 2017.
[25] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in ICMEW 2017.
[26] M. Liu, H. Liua, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," in Pattern Recognition, vol. 68, pp. 346-362, August 2017.
[27] C. Caetano, F. Bremond, and W. R. Schwartz, "Skeleton image repre-sentation for 3d action recognition based on tree structure and reference joints," in SIBGRAPI 2019.
[28] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive rein-forcement learning for skeleton-based action recognition," in CVPR 2018.
[29] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action," in CVPR 2019.
[30] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recogni-tion with directed graph neural networks," in CVPR 2019.
[31] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolution-al network for skeleton-based human action recognition by neural searching," in AAAI 2020.
[32] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recogni-tion with multi-stream adaptive graph convolutional networks," in Arxiv 2019.
[33] M. Li, S. Chen, X. Chen, Y.Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction," in Arxiv 2019.
[34] J. Gao, T. He, X. Zhou, and S. Ge, "Focusing and diffusion: Bidirection-al attentive graph convolutional networks for skeleton-based action recognition," in Arxiv 2019.
[35] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. Plan, Activity, and Intent
Recognition, 2011.
[36] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In ICCV. IEEE, 2015.
[37] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In CVPR. IEEE, 2014.

[38] H. Wang, W. Wang, and L. Wang. Hierarchical motion evolution for action recognition. In ACPR. IEEE, 2015.
[39] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In CVPR. IEEE, 2012.
[40] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In CVPR. IEEE, 2014.
[41] View Invariant Human Action Recognition Using Histograms of 3D Joints
[42] RGB-D-based Action Recognition Datasets: A Survey
[43] NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis
[44] D. Rumerhart, G. Hinton, and R. Williams, "Learning representations by back-propagation errors," Nature, vol. 323, pp. 533–536, 1986.
[45]G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.
[46] Jun Liu, Amir Shahroudy, Dong Xu, Gang Wang "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition" 2016
[47] Sijie Song, Cuiling Lan, Junliang Xing, Wen-Jun Zeng, Jiaying Liu "An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data" 2016
[48] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wen-Jun Zeng, Jianru Xue, Nanning Zheng "View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data" ICCV 2017
[49] Sijie Yan, Yuanjun Xiong, Dahua Lin "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition" 2018
[50] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, Tieniu Tan "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition" CVPR 2019
[51] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Two stream adaptive graph convolutional networks for skeleton based action recognition". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12026–12035, 2019. 1, 3, 8
[52] KeCheng, Yifan Zhang, XiangyuHe, Weihan Chen, Jian Cheng,and Hanqing Lu. "Skeleton-based action recognition with shift graph convolutional network" In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages183–192,2020.
[53] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. "Channel-wise topology refinement graph convolution for skeleton-based action recognition". In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages13359–13368,2021.2,3,5,8