مجلة جامعة دمشق للعلوم الهندسية
المجلد41 العدد (3) – 2025: 146-160

Damascus University Journal for engineering sciences
Vo 41 number (3) -2025: 146-160

تحسين كشف الأغراض ثلاثية الأبعاد من خلال المعلومات السياقية والسمات ثنائية الأبعاد

عبد العزيز ياسر النحّاس*[1] رائوف حمدان[2]

*[1]. مهندس في قسم هندسة الحواسيب والأتمتة – كلية الهندسة الميكانيكية والكهربائية – جامعة دمشق.
Abdulaziz9.alnahhas@damascusuniversity.edu.sy

[2]. مدرّس، دكتور في قسم هندسة الحواسيب والأتمتة – كلية الهندسة الميكانيكية والكهربائية – جامعة دمشق.
RaoufHamdan@Damascusuniversity.edu.sy

الملخص:

أصبحت عملية فهم المشهد في البيئة ثلاثية الأبعاد من أحد أهم الموضوعات البحثية النشطة ضمن نطاق الرؤية الحاسوبية، فبمقارنتها مع البيئة ثنائية الأبعاد ذات الصور الملونة، تحوي الصور ثلاثية الأبعاد على مجموعة ضخمة من المعلومات التي يمكن الاستفادة منها في عملية فهم المشهد وتحديد الأغراض الموجودة ضمنه، حيث تتعدد التطبيقات التي تعتمد على الفهم ثلاثي البعد للمشهد مثل تجوال الروبوتات وتطبيقات الألعاب ذات الواقع المعزّز وغيرها الكثير. إن المعلومات ثلاثية الأبعاد المحصلة بواسطة الحساسات تتصف بكونها معلومات غير بنيوية وليست مرتبة، مما يجعل عملية الاستفادة منها بالطرق التقليدية (كالشبكات العصبونية التلافيفية والشبكات العصبونية العودية) المستخدمة مع الصور ثنائية الأبعاد عند تطبيقها بصورة مباشرة أمراً صعباً.

نقدّم في هذا البحث نموذجاً جديداً (CIMVNet) للتعرف على الأغراض ضمن البيئة ثلاثية الأبعاد بالاعتماد على تعزيز معلومات الصور ثلاثية الأبعاد بمجموعة من السمات التي استُخرجت من الصور الملونة ثنائية الأبعاد، بالإضافة إلى الاستفادة من المعلومات السياقية للمشهد وارتباطات مكوناته المختلفة في تحسين دقة السمات ثلاثية الأبعاد المستخرجة. أثبت النموذج المطوّر أن الاستفادة من أكثر من مصدر للبيانات ضمن الصورة ثلاثية الأبعاد يحسّن من دقة التصنيف، فقد حقق النموذج المقترح تحسناً بالدقة مقارنة مع النتائج الخاصة بالدراسات المرجعية، بزيادة 3.04% عن أعلى دقة تم التوصّل إليها في الدراسات المرجعية السابقة.

الكلمات المفتاحية: فهم المشهد، التعرّف على الأغراض ثلاثية الأبعاد، الشبكات العصبونية التلافيفية، السمات السياقية.

تاريخ الايداع: 2023/2/20
تاريخ القبول: 2023/6/5

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

Improving 3D Object detection using contextual information & 2D features

Abdulaziz Yaser Alnahhas*[1] Raouf Hamdan[2]

*[1]. Computer Engineer in Department of Computer & Automation Engineering – Faculty of Mechanical and Electrical Engineering - Damascus University. Abdulaziz9.alnahhas@damascusuniversity.edu.sy

[2]. Dr, PHD in Department of Computer & Automation Engineering – Faculty of Mechanical and Electrical Engineering - Damascus University

RaoufHamdan@Damascusuniversity.edu.sy

Received: 20/2/2023

Accepted: 5/6/2023

Abstract:

3D object detection is becoming an active research topic in both computer vision. Compared to 2D object detection in RGB images, predicting 3D bounding boxes in real world environments captured by point clouds is more essential for many tasks, such as indoor robot navigation, robot grasping, etc. However, the unstructured data in point clouds makes the detection more challenging than in 2D. In particular, the popular convolutional neural networks (CNNs), which are highly successful in 2D object detection, are difficult to be applied to point clouds directly. In this paper, we present a new model (CIMVNet) for object recognition within the 3D environment based on enhancing the 3D image information with a set of features extracted from the 2D color images, in addition to taking advantage of the contextual information of the scene and the correlations of its various components to improve the accuracy of the 3D features. The developed model proved that taking advantage of more than one source of data within the 3D image improves the accuracy of classification. It has achieved an accuracy improvement compared to the state- of-the-art researches, which amounted to 3.04% over the highest previous reference research.

Keywords: Scene Understanding, 3d Object detection, CNNs, Contextual features.

المقدّمة:

أصبحت عملية اكتشاف الأغراض ثلاثية الأبعاد من المواضيع البحثية النشطة ضمن مجال الرؤية الحاسوبية لنظراً لأهميتها في العديد من المجالات، بالإضافة إلى انتشار أجهزة الكشف ثلاثية الأبعاد وتوافرها بشكل كبير ما ساهم في دعم هذه الأبحاث.

بمقارنة الصور ثلاثية الأبعاد مع الصور الثنائية، يمكن اعتبار صناديق الإحاطة (Bounding box) التي يتم التنبؤ بها ضمن المشهد ثلاثي الأبعاد الممثل على شكل سحابة نقاط (Point cloud) أكثر أهمية في العديد من التطبيقات مثل تجوال الروبوت والواقع المعزّز وغيرها الكثير.

إن البنية غير المهيكلة لسحابة النقاط تجعل عملية الكشف أكثر صعوبة من نظيرتها للأغراض ثنائية الأبعاد، حيث من الصعب استخدام بنى الشبكات العصبونية التلافيفية (CNNs) بشكل مباشر على الرغم من تحقيقها نتائج جيدة في عملية كشف الأغراض التقليدية في الصور الثنائية الأبعاد.

في الآونة الماضية ظهرت العديد من الدراسات التي اهتمت بشكل كبير للتعامل مع هذا التحدي. فمع ظهور شبكات معالجة النقاط ثلاثية الأبعاد العميقة مثل PointNet (Qi وآخرون،2017) [1] و++PointNet (Qi وآخرون،2017) [2]والتي شكلت مرحلة جديدة ضمن مجال معالجة سحابة النقاط، وتم اقتراح العديد من أعمال اكتشاف الأغراض ثلاثية الأبعاد القائمة على التعلم العميق مؤخراً لاكتشاف الأغراض مباشرة من السحب النقطية ثلاثية الأبعاد. اقترح بحث (Qi وآخرون، 2019) [4] شبكة VoteNet للكشف عن الأغراض ثلاثية الأبعاد من طرف إلى طرف (end-to-end) على أساس تصويت هوف (Hough voting) (Leibe وآخرون، 2004) [9].

تقوم شبكة VoteNet بتحويل إجراء (Hough voting) التقليدي إلى مشكلة عودية والتي تنفذها شبكة عميقة، حيث تقوم بأخذ عينات من عدد من النقاط الأولية من سحابة نقاط الدخل لإنشاء تصويت لمراكز الكائنات المحتملة. ثم يتم استخدام مراكز التصويت لتقدير مربعات الإحاطة ثلاثية الأبعاد (3D Bunding box). تمكن استراتيجية التصويت VoteNet من تقليل مساحة البحث بشكل كبير وتحقيق أفضل النتائج في التعامل مع العديد من مجموعات البيانات المعيارية. تفتقر شبكة VoteNet إلى مراعاة العلاقات بين الكائنات المختلفة وبين الأشياء والمشهد الذي تنتمي إليه كما أنها لا تأخذ بعين الاعتبار صورة المشهد ثنائي الأبعاد ومحتوياته في عملية الكشف، مما يحد من دقة اكتشافها.

من هنا. ظهرت العديد من المقترحات التي تساعد في تحسين أداء شبكة VoteNet وإضافة العديد من التفاصيل التي تساهم في الوصول إلى أفضل دقة ممكنة، كمحاولة الاستفادة من المعلومات السياقية للمشهد على اختلاف أنواعها، والاعتماد على السمات ثنائية الأبعاد في تصحيح نقاط المراكز وغيرها.

1. الدراسات والأبحاث ذات الصلة:

يمكن تصنيف الأبحاث والدراسات السابقة إلى مجوعة من التصنيفات من حيث الآلية المستخدمة في عملية كشف الأغراض ضمن البيئة ثلاثية الأبعاد، سنحاول التركيز على ثلاث تصنيفات رئيسية تم الاستفادة منها في هذا البحث.

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

1.2. الأبحاث التي تعتمد على المعلومات الهندسية:

لتحديد موقع الأغراض ثلاثية الأبعاد اعتماداً على المعلومات الهندسية، يعتمد أكثر الأساليب الشائعة على مفهوم مطابقة القالب (template matching) باستخدام مجموعة من نماذج CAD النقية إما بشكل مباشر مثل بحث (Litany وآخرون، 2017) [6]، أو بشكل غير مباشر من خلال السمات المستخرجة مثل بحث (Avetisyan وآخرون، النقطية كبحث (Ye وآخرون، 2018) [26]، وفهم المشهد ثلاثي الأبعاد كبحث (Zhang وآخرون، 2017) [27].

حقق بحث (Hu وآخرون، 2018) [25] نتائج ملفتة من خلال تجزئة السحب النقطية ثلاثية الأبعاد باستخدام تحليل سياق التصحيح النقطي. في حين اقترح بحث (Shi وآخرون، 2019) [23] نهج جديد يعتمد على التشفير التلقائي التكراري للتنبؤ باكتشاف الكائنات ثلاثية الأبعاد من خلال استكشاف مقدمات السياق الهرمي في تخطيط كائن ثلاثي الأبعاد مستوحاة من فكرة الاهتمام الذاتي (Self-attention) في معالجة اللغة الطبيعية (Vaswani وآخرون، 2017) [20]، تربط الأعمال الحديثة آلية الانتباه الذاتي مع التنقيب عن المعلومات السياقية لتحسين مهام فهم المشهد مثل التعرف على الصور، والتجزئة الدلالية والتعرف على سحابة النقاط وغيرها من المهام.

فيما يتعلق بمعالجة بيانات النقاط ثلاثية الأبعاد، اقترح بحث (Zhang وآخرون، 2019) [14] استخدام شبكة الانتباه لالتقاط المعلومات السياقية في نقاط ثلاثية الأبعاد. على وجه التحديد، حيث يستخدم شبكة الانتباه النقطية لتشفير السمات المحلية في واصف عام مستند على السحابة النقطية. أيضاً في بحث (Paigwar وآخرون، 2019) [13]، تم اقتراح (Attentional PointNet) للبحث في مناطق 2019) [8]. تعتمد الأساليب الحديثة على الشبكات العميقة لسحابة النقاط في سياق فهم المشهد ثلاثي الأبعاد. تعد الشبكات (Point R-CNN) (Shi وآخرون 2018) [12] و (VOTENET)(Qi وآخرون، 2019) [4] والتي أظهرت أحدث اكتشاف للأغراض ثلاثية الأبعاد في المشاهد الخارجية والداخلية هي الأقرب لنموذج المقترح ضمن هذا البحث. والجدير بالذكر أن هذه النتائج تحققت بدون استخدام أي مدخلات من صور الثنائية (RGB).

2.2. الأبحاث التي تعتمد على السمات ثنائية الأبعاد والعمق:

تحتوي قنوات العمق والألوان على معلومات مفيدة يمكن أن تساعد في اكتشاف الأغراض ثلاثية الأبعاد. تنقسم الطرق السابقة للاستفادة من هذه المعلومات إلى ثلاث فئات: ثنائية الأبعاد، وثلاثية الأبعاد، وسلسلة الميزات. النوع الأول من الطرق كالبحث (Qi وآخرون، 2018) [16] يبدأ باكتشاف الكائنات في الصورة ثنائية الأبعاد، والتي تُستخدم بعد ذلك لتوجيه فضاء البحث في الصورة ثلاثية الأبعاد. من خلال الاستناد إلى الأبعاد الثلاثية، أبحاث أخرى ركزت على دمج الميزات ثنائية الأبعاد وثلاثية الأبعاد مثل بحث (Wang وآخرون، 2019) [18].

3.2. الأبحاث التي تعتمد على المعلومات السياقية:

أظهرت دراسة (Liang وآخرون، 2013) [21] أن المعلومات السياقية لها تأثير إيجابي كبير على التجزئة الدلالية ثنائية الأبعاد واكتشاف الأغراض. منذ ذلك الحين تم استخدام المعلومات السياقية بنجاح لتحسين الأداء في العديد من المهام مثل اكتشاف الأغراض ثنائية الأبعاد كبحث (Hu وآخرون، 2018) [22]، مطابقة النقاط ثلاثية الأبعاد كبحث (Deng وآخرون، 2018) [24]، التجزئة الدلالية للسحابة

أهمية في التعرف على الكائنات من بيانات النقطة نفسها. وقد ثبت أن هذه المعلومات السياقية مفيدة في مجموعة متنوعة من مهام الرؤية الحاسوبية والتي تم ذكرها سابقاً.

في هذه الورقة، نعرض كيفية الاستفادة من المعلومات السياقية في المشاهد ثلاثية الأبعاد لتعزيز أداء اكتشاف الأغراض ثلاثية الأبعاد من السحب النقطية، إضافة إلى الاستفادة من السمات الثنائية الأبعاد واستخدامها بشكل أساسي للحد من مسح كامل سحابة نقاط الدخل.

3. النموذج المقترح CIMVNet:

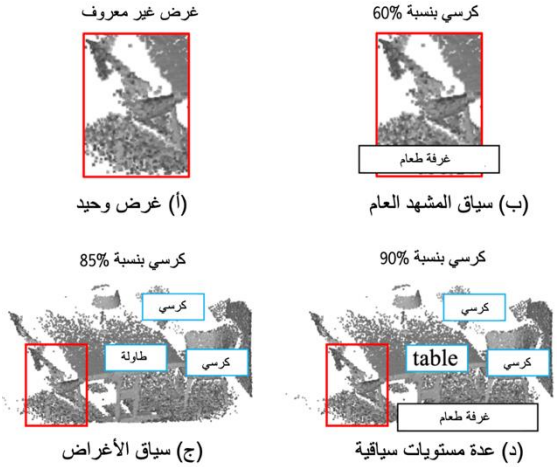
تم بناء النموذج المقترح اعتماداً على نموذج الـVoteNet (Qi وآخرون، 2019) [4] الذي يتميز بكونه يعالج دخل صورة المشهد ثلاثي الأبعاد (سحابة النقاط) بصورة مباشرة دون إجراء أية عمليات تعديل أو تحويل.

استند نموذج الـ VoteNet إلى المبدأ التقليدي في كشف الأغراض في المشاهد ثنائية الأبعاد والذي يسمى تصويت هوف (Hough voting) (Leibe وآخرون، 2004) [9] الذي يعتمد على إيجاد مراكز الأغراض ضمن الصورة ليتم بعدها تجميع النقاط حول هذا المركز وتصنيفها.

عُمِّمت هذه الفكرة على سحابة النقاط الممثلة للصورة ثلاثية الأبعاد، فبدلاً من البحث عن مركز الغرض ضمن صورة ثنائية، أصبح البحث ضمن سحابة النقاط على المحاور الثلاثة.

تبدأ عملية البحث عن مراكز الأغراض من خلال مجموعة جزئية من نقاط الدخل، يطلق على هذه المجموعة (السمات ثلاثية الأبعاد)، تُحدد السمات مباشرة من سحابة النقاط اعتماداً على نموذج Point Net++.

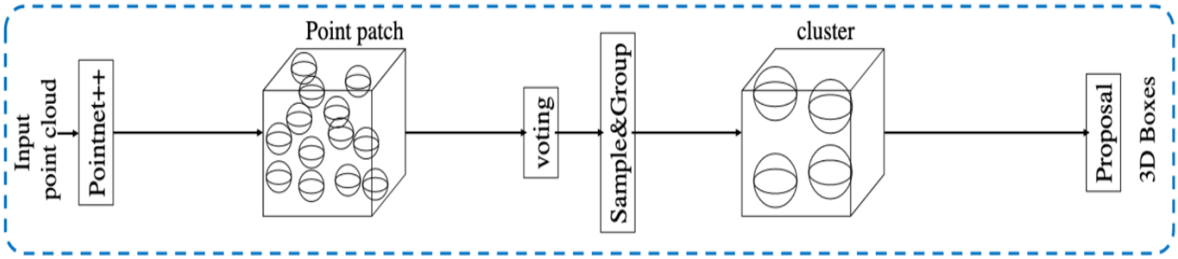
الاهتمام بدلاً من معالجة سحابة نقاط الدخل بالكامل عند اكتشاف كائنات ثلاثية الأبعاد في سحب نقطية كبيرة الحجم.



الشكل(1) أهمية المعلومات السياقية في فهم المشهد [28]

4.2. أهمية المعلومات السياقية في فهم المشهد:

يمكن تبيان أهمية المعلومات السياقية من خلال ملاحظة مثال في الشكل 1. لسحابة نقاط تم التقاطها باستخدام كاميرات العمق، والتي غالباً ما تحتوي على بيانات متداخلة ومفقودة. والذي يجعل من الصعب حتى على البشر التعرف على ماهية ومكان وجود الجسم في الشكل 1 (أ). ومع الأخذ في الاعتبار المعلومات السياقية المحيطة. في الشكل 1 (ب، د)، نجد أنه من الأسهل التعرف على أنه كرسي بالنظر إلى الكراسي المحيطة والطاولة في مشهد غرفة الطعام. في الحقيقية، يمكن أن يكون إعادة إنتاج مجموعة النقاط الممسوحة ضوئياً غامضاً عند تقديمه بشكل فردي، بسبب نقص محتوى اللون ومشاكل البيانات المفقودة. لذلك، نجادل بأن عمليات المسح المتعمق للبيئة الداخلية غالبًا ما تكون محجوبة لدرجة أن السياقات يمكن أن تلعب دورًا أكثر



الشكل (2) البنية العامّة لنموذجو [4] Votenet الذي يبين الخطوات الرئيسية لعمله (استخراج السمات، التصويت، التجميع، الفرز)

التحسينات والمساهمات البحثية التي أُضيفت على النموذج السابق للوصول إلى النموذج المقترح CIMVNet.

1.3. الاستفادة من معلومات الصور الملونة ثنائية الأبعاد:

من خلال ملاحظة خصائص بيانات سحابة النقاط (Point Cloud) وخصائص بيانات الصورة الملونة (RGB) يمكن القول إن الصفات الهندسية لبيانات سحابة النقاط غير كافية للحصول على مراكز الأجسام بشكل دقيق، لذا يمكن الاستفادة من الصور الملونة ثنائية البعد في تحسين عملية اكتشاف الأغراض في المشهد ثلاثي الأبعاد، حيث تتميز هذه الصور بدقة أعلى من صور العمق أو سحابة النقاط وتحتوي على نسج (Textures) غنية غير متوفرة في سحابة النقاط. كما يمكن أن تغطي الصور الملونة "المناطق العمياء" (blind regions) الناتجة عن أجهزة استشعار العمق والتي تحدث غالباً بسبب السطوح العاكسة ضمن المشهد.

استناداً إلى بحث (Qi وآخرون، 2020) [7] تم استخدام كاشفات ثنائية الأبعاد لتقديم مقترحات أولية في إيجاد الأغراض ثلاثية الأبعاد، وهذا يحد من عملية البحث ضمن سحابة النقاط كاملةً، حيث تم الاستفادة من أحد أكثر الكواشف الثنائية انتشاراً Fast RCNN (Girshick، 2015) [5] مع المحافظة على القدرة على اقتراح أغراض من سحابة

يمكن توضيح عمل نموذج VoteNet من خلال تسلسل الخطوات الآتية:

من أجل سحابة النقاط المدخلة ذات العدد N، تعمل الطبقة الأولى للشبكة المتمثلة بشبكة الـ PointNet++ على استخراج مجموعة السمات ثلاثية الأبعاد وهي عبارة عن مجموعة فرعية من سحابة النقاط ذات الأبعاد M، مضافاً لكل نقطة شعاع السمات الخاص بها ذات الأبعاد C.

تشكل هذه المجموعة الفرعية من النقاط بذوراً لتحديد مراكز الأغراض ضمن السحابة، كل نقطة تُصوِّت على مكان مركز الغرض الذي تنتمي إليه، ليتم بعد ذلك تحديد المراكز وفقاً للعدد الأكبر من الأصوات.

تُجمَّع كامل النقاط التي شاركت بالتصويت على مركز واحد ضمن مجموعة واحدة تعبّر عن غرض معين، لينتج مجموعات من الأغراض غير المصنّفة والمحاطة بصندوق الإحاطة (Bounding Box).

تدخل كل مجموعة إلى المصنِّف (Classifier) الذي بدوره يعطي الصنف الذي ينتمي إليه الغرض.

كما ذكرنا سابقاً، فعلى الرغم من أن هذا النموذج قد حقق نتائج ملحوظة، إلا أن هناك كثيراً من التفاصيل التي لم يأخذها بالحسبان مثل السمات ثنائية الأبعاد والمعلومات السياقية والتي تفتح الباب للعديد من الدراسات والأبحاث، وهذا ما قمنا به في هذه الرسالة، سنشرح فيما يأتي

بين أفضل ما في كلا المصدرين (الكاشف ثنائي الأبعاد الذي تم إضافته، والكاشف الرئيسي ثلاثي الأبعاد) مع تجنب عيوب كل منهما.

النقاط نفسها، وبالتالي أصبح لدينا مصدرين لاقتراح النقاط الأولية التي تساهم في علمية التصويت، ليتم بعدها الجمع

الشكل(3) بنية النموذج المقترح CIMVVNet بعد إضافة التحسين المتعلق بمعالجة الصورة الثنائية الملونة

يتم بعد ذلك (كما في نموذج VoteNet) تجميع كامل النقاط التي شاركت بالتصويت في مركز واحد ضمن مجموعة واحدة تعبر عن غرض معين، لينتج مجموعات من الأغراض غير مصنّفة محاطة بصندوق الإحاطة (Bounding Box).

تدخل كل مجموعة إلى المصنف الذي بدوره يعطي الصنف الذي ينتمي إليه الغرض.

ساهمت الإضافة المتعلقة بمعالجة الصور الثنائية الملونة بعدّة تحسينات على النموذج الأصلي تتلخص بالآتي:

- أصبح لدينا مجموعة أكبر من النقاط المساهمة في عملية التصويت ذات العدد (k) والتي أصبحت أكثر دقّة بعد حذف جزء منها (النقاط الناتجة عن الضجيج في سحابة النقاط) وإضافة أخرى (النقاط الناتجة عن المصنِّف الثنائي)
- توسيع فضاء شعاع السمات المستخرجة للنقاط لتشمل مجموعة السمات ثلاثية الأبعاد (F1) ومجموعة السمات ثنائية الأبعاد (F2) مما يعطي إمكانية أوسع للحصول على مراكز الأغراض بشكل أكثر دقة،

تتوضح مجموعة النقاط التي ستساهم في عملية التصويت بعد هذا التحسين وفق التالي: $k \times (3 + F1 + F2)$

يتمثل أحد أهم دوافع هذا التصميم في الاستفادة من الميزات الهندسية والنسيجية في الصور ثنائية الأبعاد.

يمكن تلخيص الخطوات المتعلقة بعمل النموذج بعد الإسهام المتعلق بمعالجة الصورة الملونة والمبين بالشكل 3 بالخطوات التالية:

أصبح النموذج يحتوي على فرعين تنفيذيين مستقلين، الفرع الأول هو النموذج السابق (VoteNet) المسؤول عن معالجة سحابة البيانات واستخراج النقاط الأساسية وشعاع السمات لكل منها (F1) والفرع الثاني المسؤول عن معالجة الصورة الملونة والذي يساهم أيضاً في استخراج مجموعة إضافية من السمات المتمثلة بمراكز الأجسام ثنائية الأبعاد والسمات الأصلية للبكسلات (F2).

يتم إسقاط مراكز الأجسام ثنائية الأبعاد على سحابة النقاط لتضاف على شكل مراكز زائفة إلى النقاط الأساسية المستخرجة من الفرع الأول.

تساهم كل النقاط (النقاط الأساسية والنقاط الناتجة عن الإسقاط) في عملية التصويت على مراكز الأغراض الحقيقية في البيئة ثلاثية الأبعاد.

$$A \in R^{k \times F2}$$

حيث k هي عدد النقاط الفرعية التي استُخرِجت من البيانات الخام والمقصود بها سحابة النقاط و(F2) هي طول شعاع السمات لكل نقطة. والغاية الأساسية من الوحدة الإضافية المقترحة هي الحصول على مجموعة جديدة من النقاط 'A الناتجة عن علاقة الارتباط (Correlation) بين أي نقطتين من النقاط السابقة، والموضحة بالعلاقة الآتية:

$$A' = f\big(\theta(A), \varphi(A)\big).\varrho(A)$$

حيث $\theta(.), \varphi(.), \varrho(.)$ هي ثلاث توابع تحويل مختلفة و $f(.,.)$هو تابع يرمّز التشابه بين أي موقعين من مصفوفة سمات الدخل.

وكما هو موضَّح في بحث (Hu وآخرون، 2018) [19]، تساهم هذه الارتباطات في شعاع السمات في مهام اكتشاف الأغراض، حيث يمكن الاستفادة من الشبكة غير المحلية المعممة المدمجة (CGNL) (Yue وآخرون، 2018) [3] بكونها وحدة الاهتمام (Attention module) النمطية

لنمذجة الارتباطات الغنية بين أي زوج من النقاط وأي قناة في فضاء السمات.

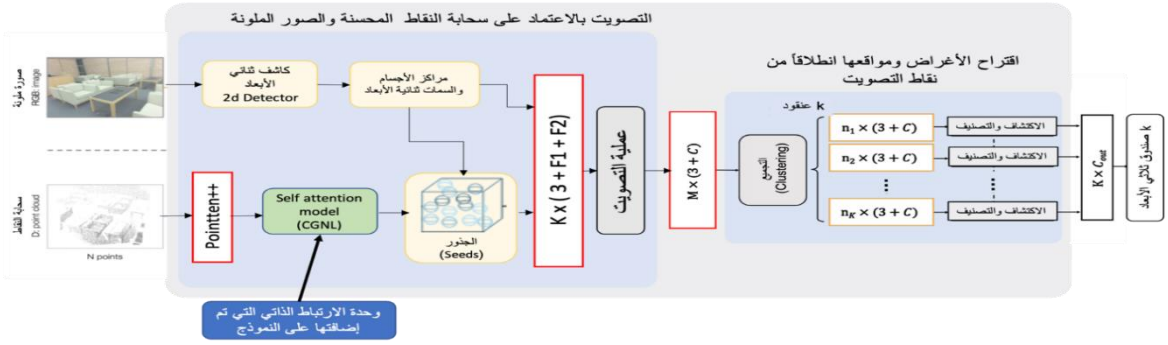
حيث كل نقطة تملك الإحداثيات المكانية (x,y,z) ومجموعة السمات التي تم توضيحها (F1, F2).

2.3. الاستفادة من المعلومات السياقية للنقاط ثلاثية الأبعاد:

كما تم التوضيح سابقاً في الفقرة 2، تساعد المعلومات السياقية الموجودة بين النقاط على تخفيف مشكلة البيانات المفقودة أو الخاطئة من خلال جمع معلومات تكميلية من نقاط مشابهة للنقاط المفقودة، لذلك تم اقتراح إضافة وحدة جديدة لالتقاط العلاقات بين النقاط وتصحيح شعاع السمات الأصلي الناتج من سحابة النقاط.

تكمن الفكرة الأساسية المستوحاة من بحث (Xie وآخرون، 2020) [10] في استخدام وحدة الاهتمام الذاتي (self-attention module) لتجميع المعلومات من جميع النقاط وإعادة بنائها مع شعاع سماتها قبل إرسالها إلى مرحلة التصويت.

بالعودة إلى النموذج الرئيسي (VoteNet)، تعمل الطبقة الأولى على استخراج السمات الرئيسية من سحابة النقاط وهي عبارة عن مجموعة جزئية من النقاط مضافاً إليها شعاع السمات، أي يمكن التعبير عن شكل الخرج على أنه



الشكل(4) المخطط العام للنموذج المطوّر (CIMVNet) بعد إضافة كل من التحسينات المتعلقة بمعالجة الصورة الثنائية الملونة و التحسينات المتعلقة بسياق النقاط ثلاثية الأبعاد

ساهمت الإضافة المتعلقة بالمعلومات السياقية بين النقاط ثلاثية الأبعاد بعدّة تحسينات تتلخّص بتصحيح فضاء

السمات ثلاثية الأبعاد (F2) ليصبح أكثر دقة بعد ملاحظة سياق كل نقطة ضمن محيطها من النقاط حيث نتج شعاع السمات الجديد (F2') المعدَّل.

يوضح الشكل 4 المخطط العام للنموذج المطوّر (CIMVNet) بعد إضافة كل من التحسينات المتعلقة بمعالجة الصورة الثنائية الملونة والتحسينات المتعلقة بسياق النقاط ثلاثية الأبعاد

4. التدريب والتقييم:

1.4. مجموعة بيانات التدريب:

هناك كثير من مجموعات البيانات التي تعنى بالأغراض والمشاهد ثلاثية الأبعاد، وقد اعتُمِد ضمن هذا البحث على أكثر مجموعة بيانات شيوعاً (Song وآخرون، 2015) (SUN_RGB-D) [11] كمعيار لتقييم النموذج المطوّر ومقارنته مع النماذج السابقة.

على عكس باقي مجموعات البيانات، فإن هذه المجموعة تحتوي بالإضافة إلى سحابة النقاط صورة ملونة ثنائية الأبعاد أحادية المشهد ،(Single-View RGB image) وهذا ما تم الاستفادة منه في الجزء المتعلق بمعالجة الصورة ضمن النموذج المطوّر.

تتألف هذه المجموعة من 10335 صورة، كل صورة تم توصيفها وتشكيل مربعات الإحاطة ثلاثية الأبعاد حول الأغراض والمصنفة إلى 37 صنفاً مختلفاً مقسمة إلى قسمين، قسم التدريب (Training Set) وقسم الاختبار (Validation Set).

تفتقر الصور ثنائية الأبعاد ضمن مجموعة البيانات إلى التوصيف (2D Labeling)، لذلك تم الاستفادة من التوصيف الذي أجري ضمن بحث (Qi وآخرون، 2020) [7] حيث سُجِّل وسُمّي كل مربع إحاطة ضمن الصور الثنائية .

أُجريت معالجة البيانات ثلاثية الأبعاد لتوليد سحابة النقاط من صور العمق (المخزنة على شكل ملفات بصيغة mat) استناداً إلى خصائص معايرة الكاميرا المرفقة (Camera Calibration) ضمن مجموعة البيانات، كما تم محاذاة كل صورة ثنائية الأبعاد مع نظيرتها (سحابة النقاط) والتي استُخدِمت للاستعلام عن مناطق الصورة المقابلة انطلاقاً من نقاط المشهد ثلاثي الأبعاد .

إن غالبية المشاهد الموجودة ضمن مجموعة البيانات مأخوذة من بيئات منزلية داخلية (indoor) ذات معلومات سياقية مرتفعة وكثيرة التداخل، ففي بعض الأحيان، يصعب على البشر أنفسهم التعرف على الأشياء في المشهد عندما يتم إعطاء سحابة نقطية ثلاثية الأبعاد فقط من دون أي معلومات لونية.

2.4. بيئة التدريب المستخدمة:

تم بناء النموذج وتدريبه واختباره بصورة كاملة على منصة (Google Colab)وهي منصة متاحة عن طريق الإنترنت، حيث تحتوي على نسختين الأولى مجانية محدودة المدّة (متاحة فقط 12 ساعة تشغيلية لكل مرّة والتي تم الاعتماد عليها) والثانية مدفوعة (مفتوحة الوقت).

تتيح هذه المنصة بيئة افتراضية بمواصفات عالية نسبياً، حيث تعمل بنظام تشغيل (Linux 18.4) وذاكرة تخزين عشوائية (12 GB RAM) ووحدة معالجة رسوميات (Tesla T4 GPU). تعمل المنصة لمدة 12 ساعة فقط ثم يتم إنهاؤها تلقائياً، ويمكن ربطها مع خدمة التخزين السحابي (Google Drive) وهذا يتيح تخزين البيانات وإعادة استرجاعها عند إنهاء المنصة وإعادة تشغيلها.

استُخدِمت لغة (Python) في إنجاز الترميز الخاص بالنموذج، كما استُخدِمت مكتبة (PyTorch) المتخصصة في بناء الشبكات العصبونية ونماذج التعلم الآلي.

3.4. بارامترات تدريب الشبكة:

تم تدريب الشبكة كاملة (end-to-end) باستخدام مُحسِّن (Adam optimizer) وحجم دفعة (Batch size) قدره 8. كما ضُبِط معدّل التعلم الأساسي (learning rate) على 0.001.

امتد تدريب الشبكة على 260 دورة تدريب (Epoch). تم تعيين خطوات تناقص معدّل التعلم (learning rate decay) عند أرقام الدورات {100، 140 ، 180} ، ومعدلات التناقص (decay rates) هي {0.1 ، 0.1 ، 0.1}. استغرقت عملية التدريب حوالي 23 ساعة.

4.4. تقييم النموذج المطوّر (CIMVNet) ومقارنته مع النماذج السابقة:

قُيِّم أداء ودقة النموذج المطوّر اعتماداً على أكثر 10 أغراض شيوعاً ضمن مجموعة البيانات (SUN RGB-D) التي وُضِحت سابقاً وذلك استناداً إلى بحث (Qi وآخرون، 2019) [4]، كما اعتُمِد مقياس (mAP) مع عتبة بمقدار 0.25 ضمن معيار (IoU)، حيث جرى مقارنة النتائج التي حصلنا عليها مع الدراسات المرجعية الأحدث في هذا السياق وفق الجدول 1

يمكن ملاحظة أن النموذج المطوّر قد تفوق في الكشف عن بعض الأغراض (مثل رف الكتب، حوض الاستحمام، الكنبة والطاولة) على باقي النماذج السابقة، في حين أن باقي الأغراض لم تحقق نتائج أفضل، وفي حال تم أخذ متوسط لكامل قيم الدقة لكل غرض من الأغراض المشمولة بالدراسة نلاحظ تفوق النموذج المطوّر بنسبة 3.04%،

تعدّ هذه النتيجة جيدة ضمن الدراسة الحالية والتي شملت إضافة المعلومات السياقية للمشهد على مستوى النقاط ومعالجة الصورة الملونة ثنائية الأبعاد. لا يمكن وصف هذه النتائج بالمميزة، إلا أنها تشكل نقطة بداية لإضافة المزيد من التحسينات التي تشمل مختلف جوانب المعلومات السياقية من المشهد وغيرها.

الجدول(1) مقارنة النتائج التي حصلنا عليها مع الدراسات المرجعية بمعيار (mAP @ 0.25)

الدراسة	تدعم الألوان	حوض استحمام (bathtub)	سرير (bed)	رف الكتب (bookshelf)	كرسي (chair)	طاولة مكتب (desk)	خزانة مطبخ (dresser)	منضدة سرير (nightstand)	كنبة (sofa)	طاولة (table)	مرحاض (toilet)	mAP
2D-driven (2017) [15]	نعم	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
PointFusion [17] (2018)	نعم	37.3	68.6	37.7	55.1	17.2	23.9	32.3	53.8	31.0	83.8	45.4
F-PointNet (2018) [16]	نعم	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	90.9	54.0
VOTENET [4] (2019)	لا	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
MLCVNet [10] (2020)	لا	79.2	85.8	31.9	75.8	26.2	31.3	61.5	66.3	50.4	98.1	59.8
IMVOTENE [7] (2020)	نعم	75.9	87.6	41.3	76.7	28.7	41.4	69.9	70.7	51.1	90.5	63.4
(النموذج الطوّر) CIMVNet	نعم	78.3	89.6	45.4	76.1	34.1	44.9	72.5	74.7	54.9	93.9	66.44

5. الخلاصة:

طُوِر في هذه البحث نموذج حاسوبي جديد (CIMVNet) للتعرف على الأغراض ضمن المشاهد ثلاثية الأبعاد، وهو البحث الأول من نوعه الذي استطاع الاستفادة من المعلومات السياقية للمشهد ثلاثي الأبعاد مع المعلومات المستخرجة من الصور ثنائية الأبعاد المرافقة للمشهد بآن معاً.

بعضاً في عملية ترميم بعض النقاط ضمن سحابة النقاط، بالإضافة إلى تحسين أسهمت الصور ثنائية الأبعاد في تصحيح بعض مواقع مراكز الأغراض في البيئة ثلاثية الأبعاد من خلال تعزيز السمات المستخرجة من سحابة النقاط بمجموعة إضافية من السمات التي تتعلق بالصور الملونة ثنائية الأبعاد. كما أسهمت المعلومات السياقية لعلاقة ارتباط النقاط مع بعضها السمات المستخرجة من خلال إعادة صياغة شعاع السمات المستخرج وإضافة تأثير كل نقطة على النقاط المحيطة.

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

إن تطبيق كل من التحسينات السابقة (الاستقادة من السمات ثنائية الأبعاد، المعلومات السياقية للمشهد) على النموذج الأولي (VoteNet) بصورة مستقلة، أوصل إلى نتائج جيدة مقارنة بالدراسات المرجعية السابقة، في حين أن دمج هذه التحسينات ضمن نموذج واحد (وهو النموذج المقترح CIMVNet) والذي استفاد من ميزات الصور الملونة والمعلومات السياقية للمشهد قد تفوق على الدراسات السابقة من خلال رفع متوسط الدقة بمقدار 3.04% والتي تعدّ حصيلة هذا البحث وخلاصته.

تم تدريب النموذج المطوّر (Sun RGB-D) على مجموعة بيانات واحدة، حيث أن مجموعات البيانات الحالية الخاصة بالمشاهد ثلاثية الأبعاد تفتقر في معظمها إلى تسمية الصور الملونة ثنائية الأبعاد للمشهد (2D image labeling) وهذا ما حدّ من إمكانية التدريب على أكثر من مجموعة، حيث تعتبر هذه النقطة هي النقطة السلبية الوحيدة ضمن هذه الرسالة.

6. آفاق مستقبلية:

من المتوقع أن النموذج الذي المطوّر في هذه الرسالة (CIMVNet) سيفتح الباب للعديد من الدراسات والأبحاث ضمن مجال التعرف على الأغراض ضمن المشهد ثلاثي الأبعاد والذي يصب أساساً في عملية فهم المشهد الكاملة.

إن المعلومات السياقية الموجودة ضمن المشهد ثلاثي الأبعاد غنية جداً وذات قيمة عالية في عملية التعرف على الأغراض، وكما ذكرنا سابقاً فإن النموذج المطور اهتم فقط بمستوى المعلومات السياقية ضمن نقاط المشهد، لذا يمكن لأي مساهمة بحثية لاحقة من تطبيق المزيد من المستويات للحصول على نتائج أفضل من حيث الدقة.

كما ذكرنا آنفاً فإن مجموعات البيانات الحالية الخاصة بالمشاهد ثلاثية الأبعاد تفتقر في معظمها إلى تسمية الصور الملونة ثنائية الأبعاد للمشهد (2D image labeling)، لذى من الممكن إعادة تدريب النموذج المطوّر على مجموعات بيانات أخرى في حال وجود التسميات لكل من سحابة النقاط والصور الثنائية ضمنها والذي يسهم في زيادة الدقة الحالية للنموذج.

التمويل: هذا البحث ممول من جامعة دمشق وفق رقم التمويل(501100020595).

References:

1- C R Qi, H Su, K Mo, and L J Guibas. (2017), PointNet: Deep learning on point sets for 3D classification and segmentation. In IEEE CVPR-2017 Conference.

2- C R Qi, Li Yi, H Su, and L J Guibas (2017), PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in neural information processing systems (30), pages 5099–5108.

3- K Yue, M Sun, Y Yuan F Zhou, E Ding, and F Xu. (2018), Compact generalized non-local network. In Advances in Neural Information Processing Systems (31), pages 6510–6519.

4- C R Qi, O Litany, K He, and L J Guibas (2019), Deep Hough voting for 3D object detection in point clouds. arXiv: 1904.09664 [cs.CV].

5- R. Girshick. (2015), Fast r-cnn. in Proceedings of the IEEE ICCV-2015.

6- O Litany, T Remez, D Freedman, L Shapira, A Bronstein, and R Gal (2017), Asist: automatic semantically in- variant scene transformation. In CVIU-2017 Conference, pages 157:284–299,

7- C R. Qi, X Chen, O Litany, L J Guibas. (2020), ImVoteNet: Boosting 3D Object Detection in Point Clouds with Image Votes. in IEEE CVPR-2020, Conference.

8- A Avetisyan, M Dahnert, A Dai, M Savva, A X Chang, and M Nießner. (2019), Scan2cad: Learning cad model alignment in rgb-d scans. In IEEE CVPR-2019 Conference.

9- B Leibe, A Leonardis, and B Schiele.(2004) Combined object categorization and segmentation with an implicit shape model. In Workshop on statistical learning in computer vision, ECCV-2004, volume 2, page 7.

10- Q Xie, Y.Lai, J Wu, Z Wang, Y Zhang, K Xu, and J Wang. (2020), MLCVNet: Multi-Level Context VoteNet for 3D Object Detection. arXiv:2004.05679v1 [cs.CV]

11- S Song, S P Lichtenberg, and J Xiao. (2015) Sun rgb-d: A rgb-d scene understanding benchmark suite, in IEEE CVPR-2015 Conference.

12- S Shi, X Wang, and H Li. (2018), Point r-cnn: 3d object proposal generation and detection from point cloud. arXiv preprint arXiv:1812.04244 [cs.CV].

13- A Paigwar, O Erkent, C Wolf, and C Laugier. (2019), Attentional PointNet for 3D-object detection in point clouds. In IEEE CVPR-2019 Workshops.

14- W Zhang and C Xiao. (2019), PCAN: 3D attention map learning using contextual information for point cloud based retrieval. In IEEE CVPR-2019, Conference, pages 12436–12445.

15- J Lahoud and B Ghanem. (2017), 2d-driven 3d object detection in rgb-d images. In IEEE CVPR-2017 pages 4622–4630.

16- C R Qi,W Liu, C Wu, H Su, and L J Guibas.(2018) Frustum pointnets for 3d object detection from rgb-d data. In IEEE CVPR-2018 Conference.

17- D Xu, D Anguelov, and A Jain. (2018), Pointfusion: Deep sensor fusion for 3d bounding box estimation. In IEEE CVPR-2018 Conference, pages 244–253.

18- C Wang, D Xu, Y Zhu, R Martın, C Lu, L Fei-Fei, and S Savarese. (2019), Densefusion: 6d object pose estimation by iterative dense fusion. In IEEE CVPR-2019 Conference, pages 3343–3352.

19- J Hu, L Shen, and G Sun. (2018), Squeeze-and-excitation networks. In IEEE CVPR-2018 conference, pages 7132–7141.

20- A Vaswani, N Shazeer, M Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. (2017), Attention is all you need. In

Advances in neural information processing systems (30), pages 5998–6008.

21- R Mottaghi, X Chen, X Liu, N G Cho, S W Lee, S Fidler, R Urtasun, and A Yuille. (2014), The role of context for object detection and semantic segmentation in the wild. In IEEE CVPR-2014 Conference, pages 891–898.

22- H Hu, J Gu, Z Zhang, J Dai, and Y Wei. (2018), Relation networks for object detection. In IEEE CVPR-2018 Conference, pages 3588–3597.

23- Y Shi, A X Chang, Z Wu, M Savva, and K Xu. (2019), Hierarchy denoising recursive autoencoders for 3D scene layout prediction. In IEEE CVPR-2019 Conference, pages 1771–1780.

24- H Deng, T Birdal, and S Ilic. (2018), Ppfnet: Global context aware local features for robust 3D point matching. In IEEE CVPR-2018 Conference, pages 195–205.

25- S M Hu, J X Cai, and Y K Lai. (2018), Semantic labeling and instance segmentation of 3D point clouds using patch context analysis and multiscale processing. In IEEE transactions on visualization and computer graphics.

26- X Ye, J Li, H Huang, L Du, and X Zhang. (2018), 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In ECCV-2018 Conference, pages 403–417.

27- Y Zhang, M Bai, P Kohli, S Izadi, and J Xiao. (2017), Deepcontext: Context-encoding neu- ral pathways for 3D holistic scene understanding. In IEEE ICCV-2017, pages 1192–1201.

28- Q Xie, Y K Lai, J Wu, Z Wang, Y Zhang, K Xu, J Wang. (2021), Vote-based 3D Object Detection with Context Modeling and SOB-3DNMS. International Journal of Computer Vision