New Researches in the Smart City
Vol 1 (Issue 4), 45-52
Nrsc.apadana.ac.ir

Apadana Institute of Higher Education

Voxelized In-Air Handwriting Recognition: Accurate 3D Pose Estimation and Recognition of Handwritten Gestures in Mid-Air Using V2V-PoseNet

Seyed Mojtaba Mirzadeh — Master of Computer Engineering, Azad University, Tehran West Branch, Tehran, Iran.

Mohammad Zare * — Department of Information Technology, Shiraz University of Technology, Shiraz, Iran.

Received: 29/January/2023 Accepted: 15/August/2023 ISSN: 2980-8936

Abstract

Most of the existing deep learning-based methods for 3D hand pose estimation from a single depth map are based on a common framework that takes a 2D depth map and directly regresses the 3D coordinates of keypoints, such as hand or human body joints, via 2D convolutional neural networks (CNNs). The first weakness of this approach is the presence of perspective distortion in the 2D depth map. While the depth map is intrinsically 3D data, many previous methods treat depth maps as 2D images that can distort the shape of the actual object through projection from 3D to 2D space. This compels the network to perform perspective distortion-invariant estimation. The second weakness of the conventional approach is that directly regressing 3D coordinates from a 2D image is a highly nonlinear mapping, which causes difficulty in the learning procedure. To overcome these weaknesses, we firstly cast the 3D hand and human pose estimation problem from a single depth map into a voxel-to-voxel prediction that uses a 3D voxelized grid and estimates the per-voxel likelihood for each keypoint. We design our model as a 3D CNN that provides accurate estimates while running in real-time.

Keywords: V2V-PoseNet, voxel-to-voxel prediction network, 3d hand pose estimation, human pose estimation, single depth map

* Corresponding Author: Mohammad.Zare@apadana.ac.ir

فصلنامه پژوهش‌های نوین در شهر هوشمند
دوره ۱، شماره ٤، تابستان ۱٤۰۲، ٤٥-٥۲
Nrsc.apadana.ac.ir

موسسه آموزش عالی آپادانا
غیر دولتی - غیر انتفاعی

تشخیص نوشتار با تشخیص حرکات دست با استفاده از واکسل‌سازی: برآورد دقیق موقعیت سه‌بعدی و تشخیص حرکات دست نوشتاری با استفاده از شبکهٔ پیش‌بینی واکسل-به-واکسل V2V-PoseNet

سیدمجتبی میرزاده | کارشناس ارشد رشتهٔ مهندسی کامپیوتر، دانشگاه آزاد واحد تهران غرب، تهران، ایران

محمد زارع * | گروه فناوری اطلاعات، دانشگاه صنعتی شیراز، شیراز، ایران

تاریخ دریافت: ۱۴۰۱/۱۱/۰۹ تاریخ پذیرش: ۱۴۰۲/۰۵/۲۴ ISSN: 2980-8936

چکیده

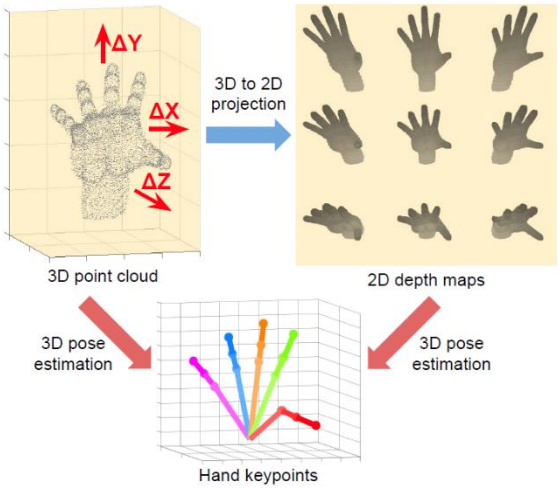
بیشتر روش‌های موجود برای تخمین ژست دست سه‌بعدی از یک نقشهٔ عمقی تک از یک چارچوب مشترک استفاده می‌کنند که با گرفتن یک نقشهٔ عمق دوبعدی، مختصات سه‌بعدی نقاط کلیدی را مستقیماً از طریق شبکه‌های عصبی کانولوشنی دوبعدی پیش‌بینی می‌کنند. اولین ضعف این رویکرد وجود انحنای پرسپکتیو در نقشهٔ عمق دوبعدی است. در حالی که نقشهٔ عمق به‌طورذاتی داده‌های سه‌بعدی است، روش‌های قبلی بسیاری از نقشه‌های عمق را به‌عنوان تصاویر دوبعدی در نظر می‌گیرند که می‌تواند شکل واقعی شیء را از طریق پراجکشن از فضای سه‌بعدی به دوبعدی تحریف کند. این مجبور به انجام تخمین مقاوم در برابر انحنای چشم‌انداز می‌شود. دومین ضعف رویکرد سنتی این است که رگرسیون مستقیم مختصات سه‌بعدی از تصویر دوبعدی، یک نقشه‌برداری بسیار غیرخطی است که موجب دشواری در روند یادگیری می‌شود. برای غلبه بر این ضعف‌ها، ابتدا مسئلهٔ تخمین ژست دست انسان سه‌بعدی از یک نقشهٔ عمقی تک را به یک پیش‌بینی وکسل‌به‌وکسل تبدیل می‌کنیم که از یک شبکهٔ وکسلی سه‌بعدی استفاده می‌کند و برای هر نقطهٔ کلیدی احتمال وکسل مربوطه را تخمین می‌زند. ما مدل خود را به‌عنوان یک شبکهٔ عصبی کانولوشنی سه‌بعدی طراحی می‌کنیم که تخمین‌های دقیقی را در حال اجرا به صورت زمان واقعی ارائه می‌دهد.

کلیدواژه‌ها: V2V-PoseNet، شبکهٔ وکسل‌به‌وکسل، Voxel-to-Voxel Prediction Network، شبکهٔ پیش‌بینی وکسل‌به‌وکسل، تخمین ژست سه‌بعدی دست، تخمین ژست انسان، نقشهٔ عمق واحد

* نویسندهٔ مسئول: Mohammad.Zare@apadana.ac.ir

مقدمه

تخمین دقیق ژست سه‌بعدی دست نیاز مهمی در تشخیص فعالیت‌ها با کاربردهای متنوع مانند تعامل انسان-کامپیوتر و واقعیت افزوده است. این مسئله در جامعهٔ بینایی ماشین برای دهه‌ها موردمطالعه قرار گرفته و با معرفی دوربین‌های عمق با قیمت پایین، به دلیل تحقیقات گسترده مجدداً موردتوجه قرار گرفته است.



شکل ۱: نمایش انحنای پرسپکتیو در تصویر عمق دوبعدی.

ابرنقطه‌های سه‌بعدی رابطهٔ یک‌به‌یک با حالت سه‌بعدی دارند؛ اما تصویر عمق دوبعدی به دلیل انحنای پرسپکتیو رابطهٔ چند به یک دارند؛ بنابراین، شبکه، مجبور است تخمینی بدون تأثیر انحنای پرسپکتیو انجام دهد. تصاویر عمق دوبعدی با ترجمهٔ ابرنقطه‌های سه‌بعدی به اندازه‌های X = -300، ۰، ۳۰۰ میلی‌متر (از چپ به راست) و Y = -300، ۰، ۳۰۰ میلی‌متر (از پایین به بالا) تولید می‌شوند. در تمام موارد، مقدار Z را برابر با ۰ میلی‌متر قرار می‌دهیم. برای نمایش، مقادیر مشابه، اندازهٔ واقعی دست انسان و پارامترهای پراجکشن دوربین در مجموعهٔ داده MSRA استفاده شده است.

اخیراً، روش‌های تمایزی قدرتمند مبتنی بر شبکه‌های عصبی کانولوشنی (CNN) در وظایف مختلف بینایی ماشین ازجمله تخمین ژست سه‌بعدی دست از یک نقشهٔ عمقی واحد بهتر از روش‌های موجود عمل می‌کنند. با این حال، این روش‌ها به دلیل انسداد شدید خود، شکل‌های بسیار پیچیدهٔ اشیاء هدف و کیفیت پایین تصاویر عمق، همچنان از تخمین نادرست رنج می‌برند.

با تحلیل روش‌های قبلی مبتنی بر یادگیری عمیق برای تخمین ژست سه‌بعدی دست از یک تصویر عمقی واحد، اکثر این روش‌ها بر پایهٔ یک چارچوب مشترک است که یک تصویر عمقی دوبعدی می‌گیرد و به‌طورمستقیم مختصات سه‌بعدی نقاط کلیدی مانند مفاصل دست را پیش‌بینی می‌کند. با این حال، ما ادعا می‌کنیم که این رویکرد دو مشکل جدی دارد. نخستین مشکل انحنای پرسپکتیو در تصاویر عمق دوبعدی است. با توجه به اینکه مقادیر پیکسل‌ها در یک نقشهٔ عمقی دوبعدی نمایانگر فواصل فیزیکی نقاط اشیاء از دوربین عمق هستند، نقشهٔ عمق درواقع داده‌های سه‌بعدی است.

با این حال، اکثر روش‌های قبلی به‌سادگی نقشه‌های عمقی را به صورت تصویر دوبعدی در نظر می‌گیرند که می‌تواند شکل واقعی یک شیء در فضای سه‌بعدی را با تصویر آن در فضای تصویر دوبعدی تحریف کند؛ بنابراین شبکه، شیءای تحریف‌شده را می‌بیند و برای انجام تخمین‌های تغییرناپذیر انحنا، بار می‌شود. ما انحناهای پرسپکتیو

تصویر عمق دوبعدی را در شکل ۱ نمایش می‌دهیم. ضعف دوم نقشه‌برداری غیرخطی بالا بین نقشهٔ عمق و مختصات سه‌بعدی است. این نقشه‌برداری غیرخطی بسیار بالا، روند یادگیری را مختل می‌کند و مانع دقیق‌ترین تخمین مختصات نقاط کلیدی توسط تامپسون و همکارانش می‌شود. این غیرخطی بالا به حقیقتی نسبت داده می‌شود که برای هر نقطهٔ کلیدی تنها یک مختصات سه‌بعدی باید از ورودی پیش‌بینی شود.

برای مقابله با این محدودیت‌ها، ما شبکهٔ پیش‌بینی وکسل‌به‌وکسل برای تخمین ژست «V2V-PoseNet» را پیشنهاد می‌کنیم. برخلاف بیشتر روش‌های قبلی، V2V-PoseNet از یک شبکهٔ وکسل‌شده به‌عنوان ورودی استفاده می‌کند و برای هر نقطهٔ کلیدی احتمال هر وکسل را تخمین می‌زند. همان‌طور که در شکل ۲ نشان داده شده است.

با تبدیل تصویر عمق دوبعدی به شکل وکسل سه‌بعدی به‌عنوان ورودی، شبکهٔ ما می‌تواند ظاهر واقعی اشیاء را بدون انحنای پرسپکتیو ببیند. همچنین، تخمین احتمال هر وکسل هر نقطه کلیدی، به شبکه کمک می‌کند تا به‌راحتی از نقشه‌برداری غیرخطی که مختصات سه‌بعدی را مستقیماً از ورودی تخمین می‌زند، کار موردنظر را یاد بگیرد. ما آزمایش جامعی را برای نشان‌دادن کارایی نمایش حجمی پیشنهادی ورودی و خروجی در تخمین ژست سه‌بعدی دست از یک نقشهٔ عمقی واحد انجام می‌دهیم. عملکرد چهار ترکیب ورودی (نقشهٔ عمق دوبعدی و شبکهٔ وکسل‌شده) و خروجی (مختصات سه‌بعدی و احتمال هر وکسل) موردمقایسه قرار می‌گیرد.

نتایج آزمایش نشان می‌دهند که پیش‌بینی وکسل‌به‌وکسل ما به روش ما امکان می‌دهد که در تقریباً تمامی مجموعه‌های دادهٔ عمومی (به‌عنوان‌مثال، سه مجموعهٔ دادهٔ ژست سه‌بعدی دست و یک مجموعهٔ دادهٔ ژست سه‌بعدی انسانی) عملکردی بهتر از روش‌های موجود ارائه دهد در حالی که به صورت زمان واقعی اجرا می‌شود.

در ادامه، ما فرض می‌کنیم که عبارت "تخمین وضعیت سه‌بعدی" به معنای مکان‌یابی نقاط کلیدی دست یا بدن انسان در فضای سه‌بعدی است.

مشارکت‌های ما را می‌توان به شرح زیر خلاصه کرد:

– ابتدا مسئلهٔ تخمین ژست سه‌بعدی از یک نقشهٔ عمقی واحد را به پیش‌بینی وکسل‌به‌وکسل تبدیل می‌کنیم. برخلاف بیشتر روش‌های قبلی که مختصات سه‌بعدی را مستقیماً از تصویر عمق دوبعدی پیش‌بینی می‌کنند، V2V-PoseNet پیشنهادی ما احتمال هر وکسل را از یک ورودی شبکهٔ وکسل‌شده تخمین می‌زند.

– ما با استفاده از تجربه‌ای ارزشمند، نمایش حجمی ورودی و خروجی را با مقایسهٔ عملکرد هر نوع ورودی (به‌عنوان‌مثال، تصویر عمق دوبعدی و شبکهٔ وکسل‌شده) و نوع خروجی (به‌عنوان‌مثال، مختصات سه‌بعدی و احتمال هر وکسل) تأیید می‌کنیم.

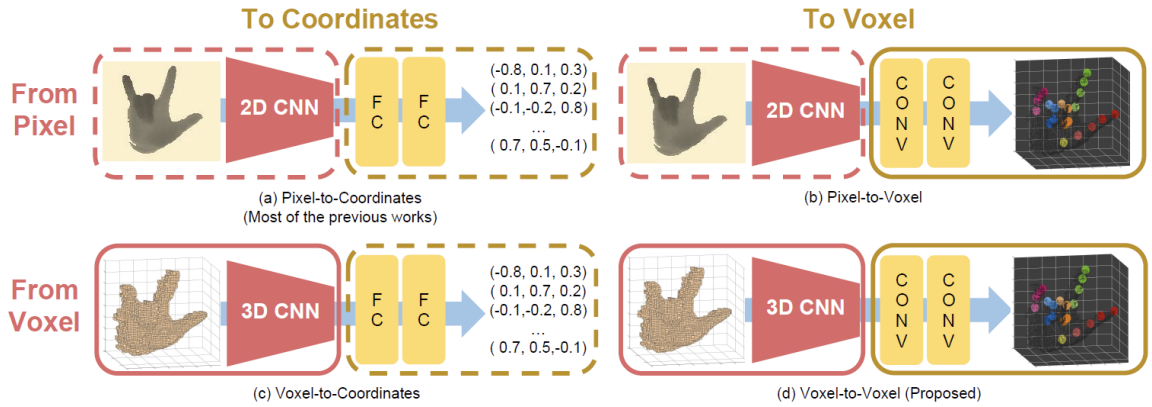
– ما آزمایش‌های گسترده‌ای با استفاده از تقریباً تمامی مجموعه‌های دادهٔ تخمین ژست سه‌بعدی موجود انجام می‌دهیم که شامل سه مجموعهٔ دادهٔ ژست سه‌بعدی دست و یک مجموعهٔ دادهٔ ژست سه‌بعدی انسانی است. نشان می‌دهیم که روش پیشنهادی نتایج بسیار دقیق‌تری نسبت به روش‌های موجود تولید می‌کند. همچنین روش پیشنهادی در چالش برآورد ژست سه‌بعدی دست بر پایهٔ فریم HANDS 2017 در رتبهٔ اول قرار گرفت.

پیشینهٔ پژوهش

تخمین ژست سه‌بعدی دست مبتنی بر عمق

روش‌های تخمین ژست دست را می‌توان به سه دستهٔ مولد، افتراقی و ترکیبی تقسیم کرد. روش‌های مولد از یک مدل پیش‌تعریف‌شدهٔ دست استفاده کرده و با کمینه‌کردن توابع هزینهٔ دست‌ساز، آن را با تصویر عمق ورودی متناسب می‌کنند. الگوریتم بهینه‌سازی ازدحام ذرات (PSO) (Sharp et al., 2015)، الگوریتم نزدیک‌ترین نقطهٔ مکرر

(ICP) (Tagliasacchi et al., 2015) و ترکیب آن‌ها (Qian et al., 2014) از الگوریتم‌های معمول برای به‌دست‌آوردن نتایج بهینه در تخمین ژست دست هستند. روش‌های مرسوم به‌صورت مستقیم مفاصل دست را در تصویر عمق ورودی مشخص می‌کنند. روش‌های مبتنی بر جنگل تصادفی، عملکرد سریع و دقیقی را ارائه می‌دهند. با این حال، آن‌ها از ویژگی‌های دست‌ساز استفاده می‌کنند و توسط روش‌های اخیر مبتنی بر شبکه‌های عصبی کانولوشنی (CNN) که قادرند ویژگی‌های مفید را به‌طورخودکار یاد بگیرند، غلبه می‌شوند. تامپسون و همکاران[1] (۲۰۱۴) برای محلی‌سازی نقاط کلیدی دست از شبکه‌های عصبی کانولوشنی استفاده کردند و برای هر مفصل دست نقشه‌های حرارتی دوبعدی تخمین زدند. گه و همکاران[2] (۲۰۱۶) این روش را با بهره‌برداری از شبکه‌های عصبی کانولوشنی چندنما گسترش دادند. گه و همکاران (۲۰۱۷) نقشهٔ عمق دوبعدی ورودی را به فرم سه‌بعدی تبدیل کرده و با استفاده از شبکه‌های عصبی کانولوشنی سه‌بعدی مستقیماً مختصات سه‌بعدی را تخمین زدند. گه و همکاران شبکه‌ای به نام شبکهٔ تجمع منطقه‌ای (Region Ensemble Network) برای دقیق‌ترین تخمین مختصات سه‌بعدی نقاط کلیدی دست پیشنهاد کردند و چن و همکاران[3] (۲۰۲۰) این شبکه را با بهبود تخمین پوز به‌طورتکراری بهبود دادند. اوبروگر و همکاران[4] (۲۰۱۷) کار قبلی خود (Oberweger et al., 2015) را با استفاده از معماری شبکه، افزایش داده و محلی‌سازی اولیهٔ دست را بهبود دادند.



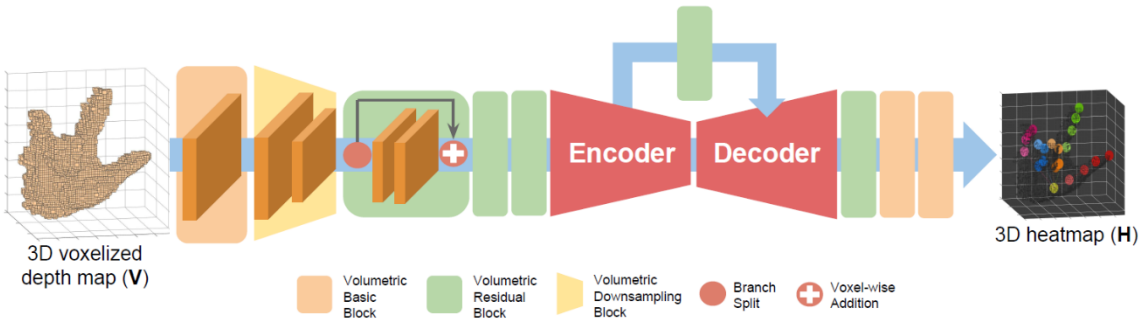
شکل ۲: ترکیب‌های مختلف ورودی و خروجی برای تخمین ژست سه‌بعدی از یک تصویر عمقی واحد.

بیشتر کارهای قبلی یک تصویر عمق دوبعدی را به‌عنوان ورودی می‌گیرند و مختصات سه‌بعدی نقاط کلیدی را مانند شکل «a» تخمین می‌زنند. به‌عکس، سیستم پیشنهادی یک شبکهٔ سه‌بعدی وکسل‌شده را به‌عنوان ورودی می‌گیرد و احتمال هر وکسل را برای هر نقطهٔ کلیدی مانند شکل «d» تخمین می‌زند. توجه کنید «b» و «d» تنها از لایه‌های کانولوشنال تشکیل شده‌اند که با یک معماری کانولوشن کاملاً شده است.

مولد و افتراقی را با استفاده از یک حلقهٔ بازخوردی آموزش دادند. ژو و همکاران[5] (۲۰۱۶) یک مدل دستی پیش‌تعریف‌شده را تعریف کردند و پارامتر مدل را به جای تخمین مستقیم مختصات سه‌بعدی استفاده کردند. یه و همکاران[6] (۲۰۱۶) از مکانیسم‌های توجه فضایی و الگوریتم بهینه‌سازی شنای ذرات سلسله‌مراتبی استفاده کردند. وان

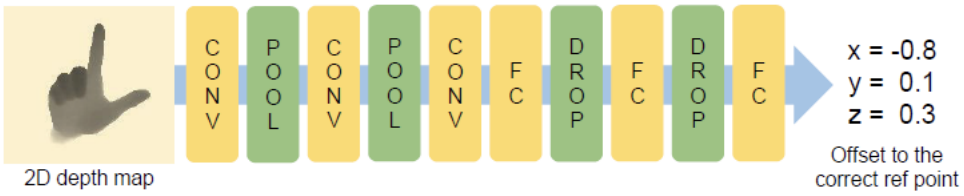
1. Tompson et al.
2. Ge et al.
3. Chen et al.
4. Oberweger et al.
5. Zhou et al.
6. Ye et al.

و همکاران[1] (۲۰۱۷) از دو مدل مولد عمیق با فضای نهان مشترک و آموزش تمیزدهنده برای تخمین پسین حالت نهفته استفاده کردند.



شکل ۳: معماری کلی V2V-PoseNet

V2V-PoseNet ورودی وکسل‌شده را می‌گیرد و با استفاده از رمزگذار و رمزگشا، احتمال هر وکسل را برای هر نقطهٔ کلیدی تخمین می‌زند. جهت ساده‌سازی تصویر، هر نقشهٔ ویژگی را بدون محور Z رسم کرده‌ایم و نقشه‌های حرارتی سه‌بعدی تمام نقاط کلیدی را در یک حجم تکی ترکیب کرده‌ایم. هر رنگ در نقشهٔ حرارتی سه‌بعدی نقاط کلیدی را در همان انگشت نشان می‌دهد.



شکل ٤: شبکهٔ بهینه‌سازی نقطهٔ مرجع.

این شبکه، تصویر عمق برش‌خورده را می‌گیرد و آفست سه‌بعدی را از نقطهٔ مرجع فعلی به مرکز مکان‌های مفصل واقعی خروجی می‌دهد.

روش

هدف مدل ما تخمین مختصات سه‌بعدی تمام نقاط کلیدی است. در ابتدا، ما تصاویر عمق دوبعدی را به فرم حجمی سه‌بعدی تبدیل می‌کنیم با بازتاب نقاط در فضای سه‌بعدی و گسسته‌سازی فضای پیوسته. پس از وکسل‌سازی تصویر عمق دوبعدی، ما از V2V-PoseNet استفاده می‌کنیم که داده‌های وکسل‌شدهٔ سه‌بعدی را به‌عنوان ورودی می‌گیرد و برای هر نقطهٔ کلیدی احتمال هر وکسل را تخمین می‌زند. موقعیت با بالاترین احتمال برای هر نقطهٔ کلیدی شناسایی شده و به مختصات واقعی جهان منحرف می‌شود که نتیجهٔ نهایی مدل ما است. شکل ۳ معماری کلی V2V-PoseNet پیشنهادی را نشان می‌دهد. در ادامه، استراتژی بهبود محلی‌سازی شیء هدف، فرآیند تولید ورودی مدل پیشنهادی V2V-PoseNet و برخی از مسائل مرتبط با رویکرد پیشنهادی را شرح می‌دهیم.

1. Wan et al.

یافته‌ها

ما پیچیدگی محاسباتی روش پیشنهادی را بررسی کردیم. زمان آموزش V2V-PoseNet برای مجموعهٔ دادهٔ ICVL دو روز، برای مجموعهٔ داده‌های NYU و MSRA 12 ساعت، برای مجموعهٔ دادهٔ چالش HANDS 2017 شش روز و برای مجموعهٔ دادهٔ ITOP سه ساعت است. زمان تست در حالت استفاده از ۱۰ مدل برای گروه دوره‌ای ۳.۵ فریم در ثانیه است، اما در محیط چند GPU می‌تواند تا ۳۵ فریم در ثانیه شتاب یابد که نشان می‌دهد قابلیت استفاده از روش پیشنهادی در برنامه‌های بلادرنگ است. مرحلهٔ محاسباتی زمان‌برتر، تولید ورودی است که شامل پالایش نقطهٔ مرجع و وکسل‌سازی نقشهٔ عمق است. این مرحله ۲۳ میلی‌ثانیه طول می‌کشد و بیشترین زمان در وکسل‌سازی صرف می‌شود. مرحلهٔ بعدی، انتقال شبکه، ۵ میلی‌ثانیه طول می‌کشد و ۰.۵ میلی‌ثانیه برای استخراج مختصات سه‌بعدی از نقشهٔ حرارتی سه‌بعدی زمان می‌برد. لازم به ذکر است که مدل ما بدون استفاده از مجموعهٔ دوره‌ای در مجموعهٔ داده‌های ICVL، NYU، MSRA و ITOP و در حالت زمان واقعی با استفاده از یک GPU واحد، عملکرد بهتری نسبت به کارهای قبلی داشت.

بحث و نتیجه‌گیری

ما یک شبکهٔ جدید و قدرتمند به نام V2VPoseNet برای تخمین ژست سه‌بعدی دست از یک نقشهٔ عمق تنها پیشنهاد کردیم. برای غلبه بر محدودیت‌های کارهای قبلی، ما نقشهٔ عمق دوبعدی را به یک نمایش وکسلی سه‌بعدی تبدیل کردیم و آن را با استفاده از مدل CNN سه‌بعدی خود پردازش کردیم. همچنین، به جای تخمین مستقیم مختصات سه‌بعدی نقاط کلیدی، ما برای هر نقطهٔ کلیدی احتمال هر وکسل را تخمین زدیم. این دو تبدیل عملکرد مدل را به‌طورقابل‌توجهی بهبود داده و باعث شد V2V-PoseNet پیشنهادی نسبت به کارهای قبلی در سه مجموعهٔ داده برای تخمین ژست سه‌بعدی دست و یک مجموعهٔ داده برای تخمین ژست سه‌بعدی انسان عملکرد بهتری داشته باشد. همچنین، با استفاده از این روش، ما توانستیم در چالش تخمین ژست سه‌بعدی دست به پیروزی برسیم. چون پیش‌بینی وکسل-وکسل برای اولین بار در تخمین ژست سه‌بعدی دست از یک نقشهٔ عمق تنها امتحان شد، امیدواریم که این کار روشی جدید برای تخمین دقیق ژست سه‌بعدی و بهبود عملکرد تشخیص دست خط را ارائه دهد.

منابع

Chen, X., Wang, G., Guo, H., & Zhang, C. (2020). Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, *395*, 138-149.

Ge, L., Liang, H., Yuan, J., & Thalmann, D. (2016). Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3593-3601).

Ge, L., Liang, H., Yuan, J., & Thalmann, D. (2017). 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1991-2000).

Oberweger, M., & Lepetit, V. (2017). Deepprior++: Improving fast and accurate 3d hand pose estimation. In *Proceedings of the IEEE international conference on computer vision Workshops* (pp. 585-594).

Oberweger, M., Wohlhart, P., & Lepetit, V. (2015). Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*.

Qian, C., Sun, X., Wei, Y., Tang, X., & Sun, J. (2014). Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1106-1113).

Sharp, T., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Kim, D., ... & Izadi, S. (2015, April). Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3633-3642).

Tagliasacchi, A., Schröder, M., Tkach, A., Bouaziz, S., Botsch, M., & Pauly, M. (2015, August). Robust articulated-icp for real-time hand tracking. In *Computer graphics forum* (Vol. 34, No. 5, pp. 101-114).

Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, *33*(5), 1-10.

Wan, C., Probst, T., Van Gool, L., & Yao, A. (2017). Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 680-689).

Ye, Q., Yuan, S., & Kim, T. K. (2016). Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14* (pp. 346-361). Springer International Publishing.

Zhou, X., Wan, Q., Zhang, W., Xue, X., & Wei, Y. (2016). Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*.

استناد به این مقاله: میرزاده، سیدمجتبی و زارع، محمد. (۱۴۰۲). تشخیص نوشتار با تشخیص حرکات دست با استفاده از واکسل‌سازی: برآورد دقیق موقعیت سه‌عدی و تشخیص حرکات دست نوشتاری با استفاده از شبکهٔ پیش‌بینی واکسل-به-واکسل V2V-PoseNet. *فصلنامه پژوهش‌های نوین در شهر هوشمند*، ۱(۴)، ۴۵-۵۲.