Evaluierung der Leistungsfähigkeit der Crowd für das Labeln von 3D-Punktwolken im Kontext von Active Learning

MICHAEL KÖLLE[1], VOLKER WALTER[1], STEFAN SCHMOHL[1] & UWE SÖRGEL[1]

Zusammenfassung: Für die automatisierte semantische Interpretation von 3D-Punktwolken bedarf es typischerweise einer großen Menge gelabelter Trainingsdaten. Diese werden oft von Experten bereitgestellt. Da es sich hierbei allerdings um einen sehr zeit- und damit kostenintensiven Prozess handelt, wird in diesem Beitrag ein Ansatz zur signifikanten Reduktion des Labelaufwands mittels Active Learning (AL) in Kombination mit Crowdsourcing vorgestellt, welcher die Einbindung eines Experten in den Labelprozess gänzlich vermeidet. Die im Rahmen des Beitrags vorgestellte Pipeline wird auf den ISPRS Vaihingen 3D Semantic Labeling Benchmark Datensatz angewandt. Hierbei kann gezeigt werden, dass mit nur 0,4% durch die Crowd gelabelter Punkte sowohl der angewandte Random-Forest-Klassifikator als auch das Sparse 3D CNN eine Overall Accuracy (OA) erreicht, die sich im Vergleich zum Training mittels des vollständig gelabelten Datensatzes um weniger als 3 Prozentpunkte unterscheidet.

1 Motivation

Gegenwärtig ist ein zunehmender Einsatz von Methoden des maschinellen Lernens, wie etwa Convolutional Neural Networks (CNNs), für verschiedene technische Anwendungen zu verzeichnen. Das Training solcher Netzwerke erfordert in der Regel allerdings enorme Mengen annotierter Daten, denen also semantische Bedeutung in Form eines sogenannten Labels zugeordnet worden ist. Im Fall von Geodaten, insbesondere im dreidimensionalen Raum, sind jedoch nur wenige solcher Datensätze vorhanden. Um entsprechende Klassifikationsansätze auch im Bereich der automatischen Interpretation von Geodaten effizient nutzen zu können, gilt es, den Labelprozess zur Bildung des Trainingsdatensatzes für diese speziellen Daten zu optimieren.
Dies kann mittels Active Learning (AL) erreicht werden (KOVASHKA et al. 2016). In WORTMAN & VAUGHAN (2018) wird der AL-Prozess als hybrides Intelligenzsystem bezeichnet, da hier eine Maschine interaktiv mit einem Menschen zusammenarbeitet, um die Stärken beider Parteien zu kombinieren. Das Ziel ist es, aus der Fülle der bislang ungelabelten Daten nur jene Untermenge zu bearbeiten, die im Sinne einer besseren Trennung der Klassen besonders hilfreich ist. Dabei identifiziert die Maschine mit Hilfe eines Machine-Learning-Algorithmus informative Instanzen, welche dann vom Menschen gelabelt werden, wodurch die Leistungsfähigkeit der Maschine zunimmt. Daher sind solche „human-in-the-loop" Systeme (BRANSON et al. 2010) ein effizientes Mittel zur signifikanten Reduktion des Labelaufwands. Dennoch müssen manuell Label bereitgestellt werden, was häufig durch Experten erfolgt. Da der Labelprozess sehr zeit- und damit kostenaufwendig ist, wächst zunehmend das Interesse, solche einfachen Labelaufgaben mittels Crowdsourcing zu lösen. Das bedeutet, der Labelprozess wird über das Internet an eine undefinierte Menge von Arbeitskräften ausgelagert.

[1] Universität Stuttgart, Institut für Photogrammetrie, Geschwister-Scholl-Str. 24D, D-70174 Stuttgart, E-Mail: [michael.koelle, volker.walter, stefan.schmohl, uwe.soergel]@ifp.uni-stuttgart.de

Mit eben dieser Methodik wurde auch bereits erfolgreich der *ImageNet*-Datensatz (DENG et al. 2009) abgeleitet. Dieser enthält über 14 Millionen annotierte 2D-Bilddaten, welche alltägliche Szenen darstellen und somit den Crowdworkern vertraut sind. Für das Labeln von 3D-Daten werden jedoch deutlich höhere Anforderungen an die Crowd gestellt, da die Interpretation von 3D-Daten, die auf einem 2-dimensionalen Bildschirm betrachtet werden, ein ausgeprägtes räumliches Vorstellungsvermögen erfordert. Die Arbeit von HERFORT et al. (2018) ist nach unserem Wissen die einzige, die sich mit 3D-Geodaten im Kontext mit Crowdsourcing auseinandersetzt.
In diesem Beitrag wird ein Ansatz zur (aus Sicht des Auftraggebers) vollautomatischen effizienten Auswahl und Annotation informativer Punkte aus einer gegebenen Punktwolke vorgestellt, wobei Crowdsourcing mit AL kombiniert wird. In dem Beitrag wird sowohl die Leistungsfähigkeit der Crowd bei der Interpretation von 3D-Daten untersucht als auch die Effizienz des AL-Prozesses aufgezeigt.

2 Methodik

2.1 Crowd-basierte Active Learning Pipeline

Die Grundidee dieser Arbeit basiert auf dem Prinzip des AL (SETTLES 2009), das auf eine Verringerung des Aufwands zur Generierung von annotierten Trainingsdaten abzielt. Dabei steht im Gegensatz zum passiven Lernen, bei dem einem Klassifikator ein bereits gelabelter Datensatz zur Verfügung gestellt wird, die Interaktion des Klassifikators mit einem Operateur während des Labeling-Prozesses im Fokus. Konkret selektiert der Klassifikator aktiv, ausgehend von einem ersten Trainingsdatensatz geringen Umfangs, welche Punkte aufgrund ihres hohen Informationsgehalts dem Trainingsdatensatz zuzuführen sind. Da diese selektierten Punkte manuell zu bearbeiten sind, ist die Maschine auf den Input eines Operateurs angewiesen. Um den kostspieligen Einsatz eines Experten nicht nur zu reduzieren, sondern gänzlich zu vermeiden, wird im Rahmen dieses Beitrags der Ansatz eines rein crowd-basierten Erfassungsprozesses verfolgt.

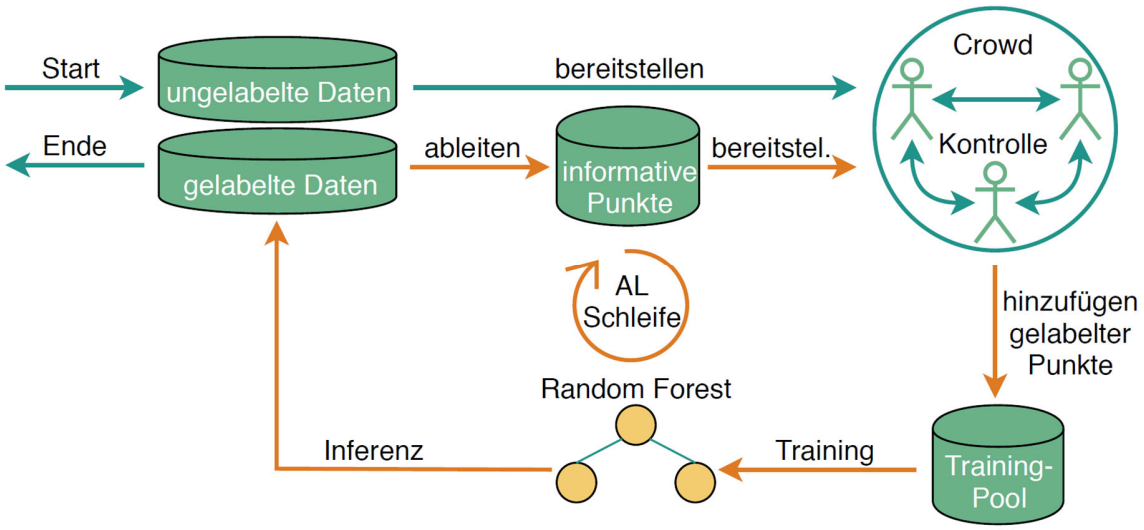

Abb. 1: Crowd-basierte AL Pipeline

Die Besonderheit liegt darin, dass, obwohl in diesem Prozess menschliche Interaktion notwendig ist, der gesamte Ablauf vollständig automatisierbar ist, da informative Instanzen automatisch durch einen Machine-Learning-Algorithmus detektiert werden. Das eigentliche Labeln durch einen Crowdworker ist aus Sicht des Auftraggebers vergleichbar mit einem Funktionsaufruf innerhalb eines Programms. Das Einstellen der Aufgabe auf einer Crowdsourcing-Plattform erfolgt automatisch unter Nutzung einer Programmierschnittstelle (API). Dasselbe gilt für die Übernahme der Ergebnisse und die Durchführung der Bezahlung. Dadurch lassen sich vollautomatisierte Prozesse realisieren, bei denen ein Teil der Aufgabe durch Menschen durchgeführt wird. Somit hat ein Auftraggeber lediglich eine Punktwolke, aus welcher der Trainingsdatensatz extrahiert werden soll, sowie das Budget zur Entlohnung der Crowdworker zur Verfügung zu stellen. Der grundsätzliche Ablaufprozess ist in Abb. 1 visualisiert.

Abb. 2: Implementiertes Webtool zur Angabe repräsentativer Vertreter der einzelnen Klassen in der voll-interaktionsfähigen Punktwolke

Die erste Aufgabe eines jeden Crowdworkers besteht darin, aus einer gegebenen Punktwolke für jede vorgegebene Klasse je einen repräsentativen Punkt zu labeln (siehe Abb. 2). Um ggf. falsch annotierte Punkte zu eliminieren, erfolgt im Rahmen einer zweiten Aufgabe eine unabhängige Kontrolle der erfassten Label. Das bedeutet, den Crowdworkern der zweiten Stufe werden jeweils selektierte 3D-Punkte präsentiert, wobei die Crowdworker zu entscheiden haben, ob diese Punkte

der korrekten Klasse zugeordnet sind. Die damit als Ergebnis erhaltenen Label tragen jedoch typischerweise einen eher geringen Informationsgehalt. Dies ist dadurch begründet, dass informative Punkte sowohl im Merkmalsraum als auch im Objektraum vornehmlich an den Entscheidungsgrenzen liegen (ERTEKIN et al. 2007). Crowdworker tendieren jedoch häufig dazu, von den Klassengrenzen weiter entfernte Punkte auszuwählen, da diese einfach und schnell einer Klasse zugeordnet werden können. Die damit erhaltenen wenigen gelabelten Punkte werden anschließend verwendet, um einen Klassifikator, in diesem Fall einen Random Forest (RF) (BREIMAN 2001), zu trainieren. Damit werden nun alle verbleibenden Punkte der Punktwolke automatisch klassifiziert, wodurch eine erste vollständige Klassifikation, allerdings von noch geringer Güte, gegeben ist.
Das zentrale Element des AL-Prozesses ist, ausgehend vom Ergebnis des RF, diejenigen Punkte zu identifizieren, die sich bis dato nur mit geringer Zuverlässigkeit klassifizieren lassen. Hierbei handelt es sich um eben jene Punkte mit hohem Informationsgehalt, welche dem Trainingspool zugefügt werden sollen. Nach SETTLES (2009) kann zur Identifikation dieser Punkte zwischen *Uncertainty Sampling* und *Query-by-Committee* unterschieden werden. Beim *Uncertainty Sampling* basiert die Wahl der informativen Instanzen auf der prädizierten a-posteriori-Wahrscheinlichkeit $P(x|c)$, dass Punkt x zu Klasse c gehört, wobei die Unsicherheit mittels Entropie bestimmt wird. Das Prinzip *Query-by-Committee* dagegen basiert auf der Evaluierung des Widerspruchs einzelner Modelle eines Ensemble-Klassifikators. Die im Rahmen dieses Beitrags genutzte *Query Function* gehört der zweiten Klasse an und wurde von ARGAMON-ENGELSON & DAGAN (1999) als Vote Entropy (VE) vorgestellt (Gleichung 1).

$$VE = -\sum_{c} \frac{\sum_e D(\boldsymbol{P}_e, c)}{N_E} \cdot \log \frac{\sum_e D(\boldsymbol{P}_e, c)}{N_E} \tag{1}$$

$$\text{wobei } D(a, c) = \begin{cases} 1, & \text{wenn } \operatorname{argmax}(a) = c \\ 0, & \text{sonst} \end{cases}$$

Dabei prädiziert jeder Klassifikator e des Ensembles E eine a-posteriori-Wahrscheinlickeit für die Klassenzugehörigkeit einer jeden Punktinstanz, welche in $\boldsymbol{P}_e$ eingeht. Basierend auf $\boldsymbol{P}_e$ wird von jedem Modell des Ensembles eine Klasse favorisiert, sodass jedes Modell für eine bestimmte Klasse stimmt. Anschließend werden die Stimmen aller Modelle aufaddiert und mit der Anzahl der Mitglieder des Ensembles N_E normiert. Für die dadurch erhaltenen Verteilungen kann die Entropie nach SHANNON (1948) bestimmt werden.
Der wesentliche Vorteil bei der Verwendung von VE als Suchfunktion gegenüber der Berechnung der gemittelten Entropie aller Klassifikatoren des Ensembles liegt darin, dass Punkte, deren maximale a-posteriori-Wahrscheinlichkeit gering ausfällt, nicht notwendigerweise selektiert werden, solange die Ensemble-Mitglieder für dieselbe Klasse stimmen. Im Rahmen dieser Arbeit ist der Ensemble-Klassifikator durch den in KÖLLE et al. (2019) vorgestellen RF-Klassifikator gegeben, wobei sowohl geometrische als auch radiometrische Merkmale genutzt werden.
Auf Grundlage der bestimmten VE für jeden 3D-Punkt, werden dann diejenigen Punkte mit der höchsten VE ausgewählt und den Crowdworkern im Rahmen einer dritten Crowdaufgabe zum Labeln angeboten und per Mehrheitsentscheid überprüft. Nach entsprechender Ergänzung des Trainingsdatensatzes um diese nun gelabelten Punkte kann jeweils ein erneutes Trainieren des RF

und ein automatisches Klassifizieren der verbleibenden Punktwolke erfolgen. Es handelt sich also um einen iterativen Prozess, in welchem der Trainingsdatensatz schrittweise aus Punkten mit hohem Informationsgehalt aufgebaut wird mit dem Ziel, die Performance des Klassifikators inkrementell zu erhöhen. Die Iteration wird solange fortgesetzt, bis das Ergebnis der gewünschten Qualität entspricht oder das Label-Budget erschöpft ist.

2.2 Datengrundlage

Die in Kapitel 2.1 vorgestellte Pipeline wird auf den *ISPRS Vaihingen 3D Semantic Labeling Benchmark* Datensatz (V3D) (NIEMEYER et al. 2014) angewandt (Klassen: *Powerline*, *Low Vegetation*, *Impervious Surface*, *Car*, *Fence/Hedge*, *Roof*, *Façade*, *Shrub*, *Tree*). Diese mittels Airborne Laserscanning (ALS) erfasste Punktwolke bildet ein für westliche Länder typisches Vorstadtgebiet ab. Die Punktdichte dieses Datensatzes liegt bei 4-8 Punkten/m^2. Um den Crowdworkern das Erkennen bekannter Objekte und Strukturen zu erleichtern, wird die Punktwolke zusätzlich mithilfe des in CRAMER (2010) abgeleiteten Orthophotos durch orthogonale Projektion koloriert (siehe Abb. 3). Da jedoch die Bilddaten (6. August 2008) und die ALS-Daten (21. August 2008) nicht zeitgleich erfasst wurden, ergeben sich insbesondere bei dynamischen Objekten wie Fahrzeugen Artefakte.

Abb. 3: Mittels Orthophoto kolorierter V3D-Trainingsdatensatz

3 Ergebnisse

Im Rahmen dieses Kapitels werden die Ergebnisse bei Anwendung der vorgestellten Pipeline auf den V3D-Datensatz diskutiert. Um abschätzen zu können, inwieweit das manuelle Labeln von 3D-Punkten crowd-basiert erfolgen kann und um generalisierbare Ergebnisse zu erhalten, ist es entscheidend, dass alle Experimente mit unabhängigen Crowdworkern durchgeführt werden, die nicht explizit instruiert werden. Auf eine solche unabhängige Crowd kann unter Nutzung der Plattform

Microworkers zurückgegriffen werden, welche in HIRTH et al. (2011) detailliert analysiert wird. Laut WEBLABCENTER INC. (2019) kann über diese Plattform auf ca. 1,4 Millionen Crowdworker zurückgegriffen werden, welche in verschiedenen Kategorien, wie zum Beispiel *Top Perfomers*, *All EU Workers* oder *Top EU Workers* organisiert sind.
Um eine Crowd-Kampagne auf *Microworkers* zu starten, müssen neben der exakten Aufgabenbeschreibung die anfallenden Gebühren übermittelt werden. Anschließend wird diese Kampagne seitens *Microworkers* freigeschaltet, sodass alle Crowdworker der spezifizierten Gruppe an dieser Kampagne mitarbeiten können. Im Rahmen der Anleitung auf der Plattform wird die URL zu den implementieren Webtools bereitgestellt. Nach Beendigung einer Aufgabe kann dem jeweiligen Crowdworker ein Code ausgestellt werden, welcher diesem als Nachweis zum Erhalt des Honorars dient.
Bei der Evaluation der Ergebnisse wird zum einen die erreichte Labelgenauigkeit der Crowdworker bei der manuellen Annotation von 3D-Punktwolken untersucht. Zum anderen wird die Performance des RF-Klassifikators über die gesamte Pipeline hinweg analysiert, wobei diese maßgeblich von der Labelgüte der Crowd bestimmt wird, da hier die Maschine, repräsentiert durch den RF, vom Menschen, genauer der Crowd, lernt.

3.1 Labelgenauigkeit der Crowd

Im Gesamten wurden 10 Crowd-Kampagnen (2 für die Initialisierung und 8 AL-Iterationsschritte), bestehend aus 1016 Aufgaben (100 freie Labelaufgaben, 100 Kontrollaufgaben und 102 Labelaufgaben pro Iterationsschritt) für das Labeln bzw. Kontrollieren von 3348 3D-Punkten (100·9 Klassen + Batchgröße von 306·8 Iterationsschritten) durchgeführt. Jede Aufgabe wurde an die *Microworkers*-Gruppe *Top Performers,* welcher mehr als 2000 aktive Worker angehören, gestellt und, je nach Aufgabentyp, mit 0,10$ bis 0,15$ vergütet.

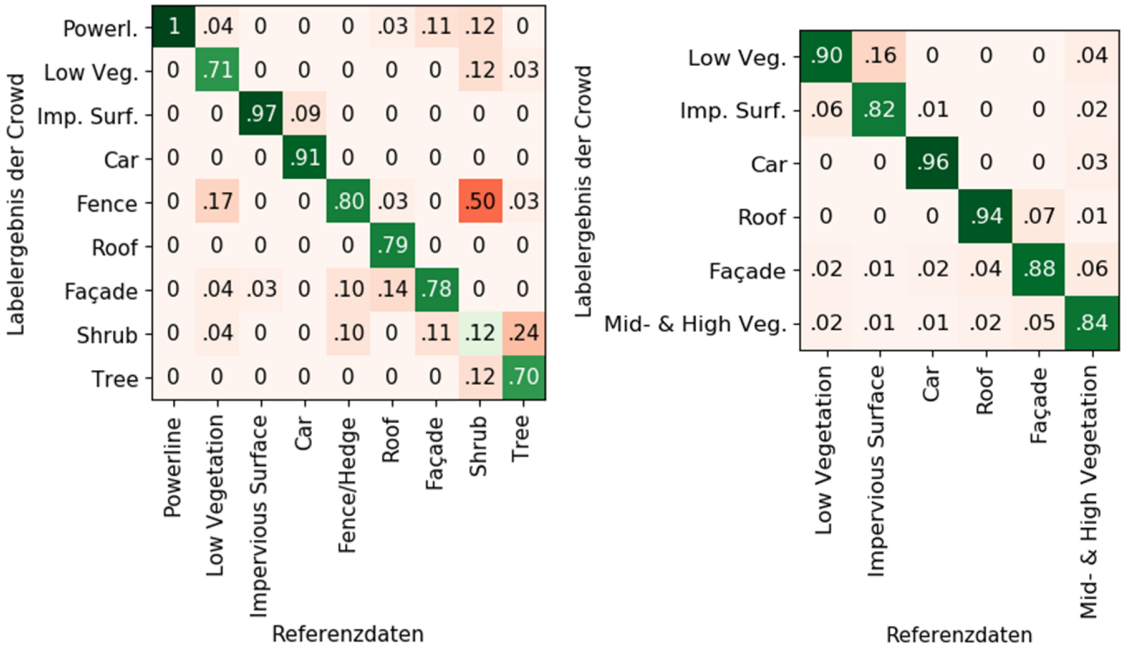

Abb. 4: Konfusionsmatrizen für die Labelgenauigkeit der Crowd für die Initialisierung (*links*) und über alle Iterationsschritte hinweg (*rechts*)

Die Analyse der Ergebnisse der von der Crowd im Rahmen der ersten Aufgabe frei gewählten und gelabelten Punkte ergibt erwartungsgemäß zunächst eine geringe Overall Accuracy (OA) von 56,44%, da hierbei die unkontrollierten Ergebnisse einzelner Crowdworker eingehen. Darüber hinaus ist zu berücksichtigen, dass im Zusammenhang mit dem im Rahmen dieses Beitrags angewandten bezahlten Crowdsourcings der hauptsächliche Anreiz der Worker im Honorar liegt und daher Crowdworker oft versuchen, entsprechende Aufgaben so schnell wie möglich abzuschließen, um ihren Gewinn zu maximieren (WORTMAN VAUGHAN 2018). Daher sind effiziente Kontrollmechanismen unerlässlich, um qualitativ hochwertige Ergebnisse zu erhalten. Mittels der in Kapitel 2.1 diskutierten Kontrollaufgabe, welche der Crowd gestellt wird, kann die OA bereits auf 71.79% erhöht werden. Die zugehörige Konfusionsmatrix ist in Abb. 4 (*links*) dargestellt.
Es zeigt sich, dass trotz der unabhängigen Kontrolle durch die zweite Gruppe von Crowdworkern insbesondere Vegetationsklassen (*Fence/Hedge*, *Shrub*, *Tree*) von den Crowdworkern verwechselt werden. Da die subjektiv beeinflusste definitionsabhängige Trennung dieser Klassen selbst für Experten eine Herausforderung darstellt, werden diese zur Klasse *Mid- & High Vegetation* fusioniert. Die fälschliche Zuordnung von Punkten zur Klasse *Powerline* ist durch den typischerweise geringen Anteil von Vertretern dieser Klasse in ALS-Punktwolken begründet. Daher werden von den Crowdworkern oft zufällig Punkte ausgewählt, um die Aufgabe möglichst schnell beenden zu können und die Bezahlung zu erhalten. Auch kann beobachtet werden, dass andere linienförmig angeordnete Punkte, wie sie etwa im V3D-Datensatz entlang einiger Fassaden auftreten (siehe Abb. 5, *rechts)*, fälschlicherweise als *Powerline* gelabelt werden. Aufgrund der diskutierten Schwierigkeiten wird die Klasse *Powerline* für die weitere Auswertung mit der Klasse *Roof* vereinigt. Mittels dieser reduzierten Klassenanzahl kann bereits eine OA von 88,59% erreicht werden, sodass ein geeigneter initialer Datensatz gestellt werden kann. Dementsprechend liegt der Fokus der Pipeline auf Klassen, welche sich für eine Erfassung durch die Crowd als geeignet herausgestellt haben.

Abb. 5: Fehlklassifikationen durch die Crowdworker (gelb). Diese sind teils auf Verdeckungen (*links*) zurückzuführen, teilweise sind diese aber auch definitionsabhängig (*rechts*)

Werden die von der Crowd erhaltenen Label über alle Iterationsschritte hinweg betrachtet, wird eine sehr ähnliche OA von 88,87% erhalten (Konfusionsmatrix in Abb. 4, *rechts*). Beispielhaft werden für weiterhin auftretende Klassenverwechslungen in Abb. 5 zwei Szenarien dargestellt, anhand derer die fehlerhaften Zuordnungen erklärbar sind. Beide Beispiele zeigen, dass die in Kapitel 2.1 vorgestellte *Query Function* zielführend Punkte, die sowohl im Merkmalsraum als

auch im Objektraum jeweils auf der Entscheidungsgrenze liegen, selektiert. Jedoch stellt die korrekte Klassenzuweisung dieser Punkte für Crowdworker offensichtlich oft eine Herausforderung dar. Im ersten Fall wurde ein der Klasse *Impervious Surface* angehörender Punkt,
welcher von der benachbarten Hecke leicht verdeckt wird, fälschlicherweise der Klasse *Low Vegetation* zugeordnet. Dies zeigt, dass für die Crowd solche Verdeckungen nur schwerlich zu interpretieren sind. Das zweite Beispiel veranschaulicht, dass Klassenzuordnungen oft interpretationsabhängig sind, was anhand der im Datensatz enthaltenen Wohnblöcke mit Flachdächern veranschaulicht werden kann. Punkte, welche direkt auf der Kante zwischen den Klassen *Roof* und *Façade* liegen, werden hier häufig, je nach eigener Interpretation, der jeweils anderen Klasse zugeordnet.

3.2 Zeit- und Kostenaufwand der Crowd-Kampagnen

Neben den von der Crowd erhaltenen Ergebnissen wird zusätzlich der Zeitaufwand für die Bearbeitung einer einzelnen Aufgabe sowie einer vollständigen Crowd-Kampagne diskutiert, wobei dieser nach WALTER & SÖRGEL (2018) von der Bezahlung abhängt. Anhand Tab. 1 wird bezüglich dieser beiden Aspekte eine Übersicht für die im Rahmen dieses Beitrags durchgeführten Crowd-Kampagnen gegeben.

Tab. 1: Zeit- und Kostenaufwand für die durchgeführten Crowd-Kampagnen.

Kampagne	mittlere Dauer/ Aufgabe [min]	Dauer/ Kampagne [h]	Bezahlung/ Aufgabe [$]
Freie Punktselektion	6,42	23,60	0,10
Kontrolle	4,70	17,20	0,12
Iteration Ø	2,53	14,80	0,15

Ausgehend von der mittleren Dauer, die für die Bearbeitung der einzelnen Aufgaben anfällt, zeigt sich, dass die Selektion beliebiger Punkte als Vertreter der jeweiligen Klasse erwartungsgemäß die meiste Zeit in Anspruch nimmt, da hierfür eine Navigation innerhalb der Punktwolke notwendig ist. Des Weiteren benötigen die Crowdworker mehr Zeit für das Kontrollieren gegebener Label als für das Labeln bestimmter Punkte innerhalb der AL-Iterationsschritte. Zwar erfordern beide Aufgaben eine eigenständige Interpretation der markierten Punkte, jedoch ist bei der Kontrollaufgabe zudem ein Vergleich mit der angegebenen Klasse erforderlich. In Bezug auf das Labeln informativer Instanzen im Rahmen der AL-Iteration ergibt sich für die einzelnen Aufgaben eine Standardabweichung von 0,2 min, was darauf hinweist, dass bezüglich des Zeitaufwands die tatsächliche Klassenzugehörigkeit der einzelnen Punkte keinen Einfluss hat. Im Mittel entsteht pro Iterationsschritt für das manuelle Labeln ein Zeitaufwand von 14,80h.
Im Rahmen der Crowd-Kampagnen wurden an die Crowdworker insgesamt 144,40$ (100·0,10$ + 100·0,12$ + 102·8·0,15$) für das Bereitstellen der Label zum Aufbau des Trainingsdatensatzes ausbezahlt. Dabei ist zu berücksichtigen, dass die Anzahl der notwendigen Label mit zunehmender räumlicher Ausdehnung der Punktwolken nicht zwangsweise ansteigt. Zwar kann davon ausgegangen werden, dass eine Punktwolke, welche einen größeren räumlichen Bereich abdeckt, auch Punkte enthält, deren Merkmale sich von denjenigen, wie sie in einem enger begrenzten Gebiet auftreten, unterscheiden, sodass weitere informative Punkte im Datensatz enthalten sind. Jedoch

wird vor allem die Redundanz von Punkten mit ähnlichen Merkmalen erhöht. Beispielsweise entsteht bei der Klasse *Roof* ein zusätzlicher Labelaufwand lediglich für Punkte, welche auf Dachformen liegen, wie sie evtl. nur in einem größeren Gebiet auftreten. Jedoch überwiegen auch hier, analog zu einem kleineren Ausschnitt, insbesondere Flach- und Giebeldächer.

3.3 Performance der Active Learning Pipeline

Auf Grundlage der von der Crowd generierten Label (vgl. Kapitel 3.1) kann der RF-Klassifikator im Rahmen des AL-Prozesses pro Iterationsschritt neu trainiert (vgl. Kapitel 2.1) und dessen Ergebnis analysiert werden. Hierfür wird das bisher nicht verwendete Testgebiet des V3D-Datensatzes herangezogen. In Tab. 2 werden anhand der OA die Genauigkeiten für die Initialisierung der AL Pipeline sowie die 8 durchgeführten Iterationsschritte dargestellt.

Tab. 2: Erzielte OA des auf den Testdatensatz von V3D angewandten RF in Abhängigkeit des durch jeden Iterationsschritt erweiterten Trainingspools.

	Iterationsschritt								
	0	1	2	3	4	5	6	7	8
OA [%]	52,19	61,38	72,32	71,79	77,93	81,21	84,48	85,62	85,82

Das schrittweise Hinzunehmen informativer Punkte führt anfangs zu einem rapiden Anstieg der OA, welcher dann bis zu Iterationsschritt 8 mehr und mehr abflacht. Insgesamt konnte diese, beginnend bei einem geringen Wert von 52,19% nach der Initialisierung, durch das selektive Erweitern des Trainingsdatensatzes um informative Instanzen auf 85,82% gesteigert werden.
Um neben der OA, welche wesentlich durch die Prädiktionsgenauigkeit in Bezug auf stark repräsentierte Klassen bestimmt wird, auch die Fähigkeit des AL-Prozesses zur Selektion und daraus resultierender Prädiktionsgüte für unterrepräsentierte Klassen zu bewerten, wird in Abb. 6 die relative Klassenhäufigkeit aller Punkte des V3D-Trainingsdatensatzes (siehe Abb. 6 *links*) mit derjenigen der im Rahmen der AL Pipeline selektierten Punkte (siehe Abb. 6 *rechts*) verglichen. Es zeigt sich, dass trotz des signifikant geringeren Punktanteils der Klassen *Car* und *Façade* im V3D-Datensatz, mittels des vorgestellten Ansatzes nahezu eine Gleichverteilung der Klassen im schrittweise aufgebauten Trainingsdatensatz erreicht werden kann, wobei pro Iterationsschritt 306 Punkte (Anzahl experimentell bestimmt) ausgewählt werden. Die erreichte Gleichverteilung der Klassen ist darauf zurückzuführen, dass die *Query Function*, ausgehend vom initialen Trainingsdatensatz, jeweils aus dem gesamten V3D-Trainingsdatensatz diejenigen Punkte selektiert, deren

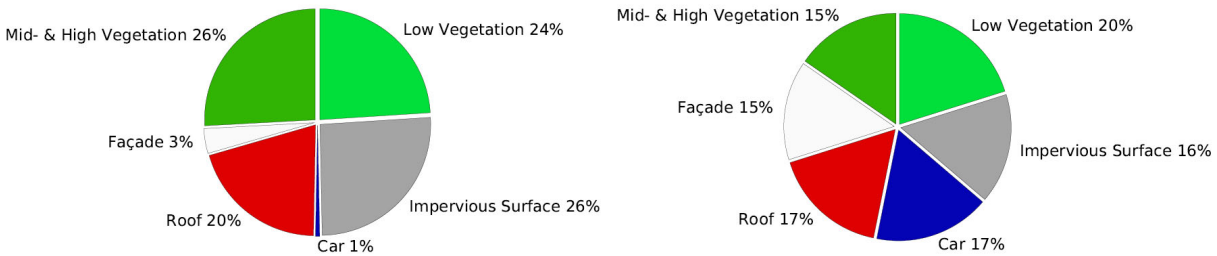

Abb. 6: Relative Klassenhäufigkeit aller Punkte des V3D Trainingsdatensatzes (*links*) und der im Rahmen der crowd-basierten AL Pipeline ausgewählten Punkte (*rechts*).

Merkmale für den Klassifikator bis dato unbekannt sind und damit den größten Widerspruch/Informationsgehalt aufweisen. Deshalb werden auch Punkte unterrepräsentierter Klassen in den Trainingsdatensatz aufgenommen, jedoch nur solange deren Merkmalsvektoren ausreichend unterschiedlich sind. Dadurch kann ein Quasi-Oversampling vermieden werden. Diese Möglichkeit, AL zur Erzeugung von Datensätzen nahezu gleichverteilter Klassenzugehörigkeit zu nutzen, wurde auch bereits von ATTENBERG & ERTEKIN (2013) beobachtet. Dadurch erübrigt sich die explizite Anpassung der Klassifikatoren an die jeweilige Klassenverteilung, beispielsweise über die Wahl einer Kostenfunktion (HE & GARCIA 2009), um weniger stark vertretene Klassen ausreichend berücksichtigen zu können. Dies stellt insbesondere für Anwendungen im Bereich der Fernerkundung einen entscheidenden Vorteil dar, da den meisten ALS-Punktwolken eine unausgewogene Klassenverteilung inhärent ist.
Um die Effizienz des vorgestellten Verfahrens aufzuzeigen, wird die Prädiktionsgenauigkeit des RF für die einzelnen Klassen bei Verwendung des in diesem Beitrag abgeleiteten Trainingsdatensatzes anhand der F1-Scores (GOUTTE & GAUSSIER 2005) jeweils derjenigen gegenübergestellt, welche sich bei Einsatz des gesamten V3D-Trainingsdatensatzes (reduzierte Klassenanzahl) ergibt (siehe Tab. 3). Es zeigt sich, dass der verwendete Ansatz nach nur 8 Iterationsschritten eine um weniger als 3 Prozentpunkte geringere OA hervorbringt.
Da der RF-Klassifikator bereits bei der Ableitung des Trainingsdatensatzes involviert war, wird zur Demonstration der allgemeinen Anwendbarkeit des erzeugten Trainingsdatensatzes zusätzlich gezeigt, dass dieser auch für das Training der typischerweise auf einen großen Trainingspool angewiesenen CNNs geeignet ist. Hierfür wird das in SCHMOHL & SÖRGEL (2019) genutzte 3D-Submanifold Sparse Convolutional Network (SSCN) anhand des abgeleiteten Trainingsdatensatzes trainiert. Tab. 3 zeigt auf, dass mit diesem Klassifikator im Vergleich zum RF sehr ähnliche Ergebnisse erzielt werden. Dadurch wird deutlich, dass ein kleinerer, mannigfaltig aufgebauter Trainingsdatensatz, welcher sich aus informativen Instanzen zusammensetzt, einem großen redundanten Trainingsdatensatz vorzuziehen ist.

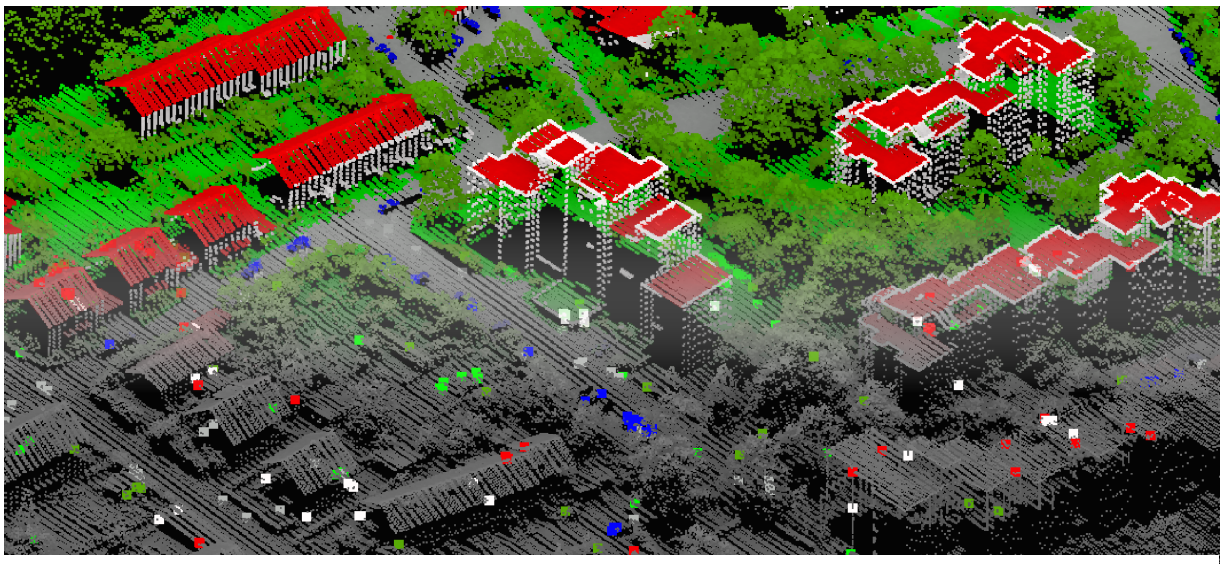

Abb. 7: Mittels der vorgestellten Pipeline kann die durch einen Experten vollständig gelabelte Trainingspunktwolke (*obere Bildhälfte*) durch das Labeln weniger informativer Punkte durch die Crowd ersetzt werden (*untere Bildhälfte*).

Hervorzuheben ist, dass diese Genauigkeiten ohne die Verwendung der im Rahmen des V3D-Datensatzes bereitgestellten Trainingsdaten erreicht wurden, indem lediglich 0,4% der Punkte des verfügbaren Trainingsdatensatzes durch die Crowd gelabelt wurden (siehe Abb. 7). Um dieses Klassifikationsergebnis zu erreichen, musste der Crowd lediglich die ungelabelte Trainingspunktwolke und ein finanzielles Budget von ca. 145$ für die Entlohnung der Worker zur Verfügung gestellt werden.

Tab. 3: Vergleich der Klassifikationsergebnisse für den Testdatensatz von V3D bei Verwendung des V3D-Trainingsdatensatzes vs. Verwendung der durch die crowd-basierte AL Pipeline abgeleiteten Trainingsdaten.

		F1-Score [%]						
Klassifikator	Trainings-datensatz	Low Ve-getation	Impervious Surface	Car	Roof	Façade	Mid-and High Vegetation	OA [%]
RF	V3D	82,55	91,84	69,85	95,01	62,67	86,15	88,42
	CB-AL	79,91	86,69	68,23	93,86	61,57	85,77	85,82
SSCN	V3D	82,48	91,16	75,15	94,89	61,53	87,29	88,39
	CB-AL	80,00	88,14	75,20	91,08	57,34	84,72	85,43

4 Zusammenfassung und Ausblick

Im Rahmen dieses Beitrags wurde gezeigt, dass ein kompakter, aus informativen Instanzen aufgebauter Trainingsdatensatz, mittels eines vollautomatisierbaren Frameworks abgeleitet werden kann, indem Crowdsourcing mit Machine-Learning-Techniken kombiniert wird. Dieser hybride Prozess umfasst die automatische Selektion eben dieser informativen Punkte, welche dann manuell durch Crowdworker gelabelt werden. Durch die Nutzung eines solchen Datensatzes für das Training der Klassifikatoren konnte der notwendige Labelaufwand, bei annähernd gleicher Prädiktionsgenauigkeit im Vergleich zum vollständig gelabelten V3D-Trainingsdatensatz, signifikant reduziert werden.

Des Weiteren haben die Experimente ergeben, dass die Crowd nicht in der Lage ist, beliebige Anforderungen eines Auftraggebers zu erfüllen. Insbesondere die subjektiv beeinflusste Trennung von Vegetationsklassen, welche selbst für Experten eine Herausforderung darstellt, ist für die Crowdworker nur schwer zu bewältigen. Während die Punktdichte des genutzten V3D- Datensatzes im Bereich der Landesvermessung häufig realisiert wird, kann die räumliche Auflösung in UAV-Befliegungen erheblich gesteigert werden (CRAMER at al. 2018). Es ist davon auszugehen, dass solche hochaufgelösten Punktwolken für Crowdworker deutlich einfacher zu labeln sind und die Möglichkeit bieten, detaillierte Strukturen, wie etwa Fassaden oder Dachinformationen, zu extrahieren.

5 Literaturverzeichnis

ARGAMON-ENGELSON, S. & DAGAN, I., 1999: Committee-Based Sample Selection For Probabilistic Classifiers. Artif. Intell. Res. **11**, 335-360.

ATTENBERG, J. & ERTEKIN, S., 2013: Class Imbalance and Active Learning. Imbalanced Learning, John Wiley & Sons Ltd., 101-149.

BRANSON, S., WAH, C, SCHROFF, F., BABENKO, B., WELINDER, P., PERONA, P. & BELONGIE, S., 2010: Visual Recognition with Humans in the Loop. ECCV 2010, 438-451.

BREIMAN, L., 2001: Random Forests. Mach. Learn. **45**(1), 5-32.

CRAMER, M., 2010: The DGPF-Test on Digital Airborne Camera Evaluation – Overview and Test Design. Photogrammetrie-Fernerkundung-Geoinformation 2010(2), 73-82.

CRAMER, M., HAALA, N., LAUPHEIMER, D., MANDLBURGER, G. & HAVEL, P., 2018: Ultra-High Precision UAV-Based LiDAR and Dense Image Matching. ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-1, 115-120.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & LI, F. F., 2009: ImageNet: A Large-Scale Hierarchical Image Database. CVPR 2009, 248-255.

ERTEKIN, S., HUANG, J., BOTTOU, L. & GILES, L., 2007: Learning on the Border: Active Learning in Imbalanced Data Classification. Proc. 16th ACM Conf. inform. knowledge management, 127-136.

GOUTTE, C. & GAUSSIER, E., 2005: A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 345-359.

HE, H. & GARCIA, E., 2009: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263-1284.

HERFORT, B., HÖFLE, B. & KLONNER, C., 2018: 3D micro-mapping: Towards assessing the quality of crowdsourcing to support 3D point cloud analysis. ISPRS Jour. Photogramm. Remote Sens. **137**, 73-83.

HIRTH, M., HOßFELD, T. & TRAN-GIA, P., 2011: Anatomy of a Crowdsourcing Platform – Using the Example of Microworkers.com. Proc. IMIS'11, IEEE Computer Society, Washington D.C., USA, 322-329.

KÖLLE, M., LAUPHEIMER, D. & HAALA, N., 2019: Klassifikation hochaufgelöster LiDAR- und MVS-Punktwolken zu Monitoringzwecken. Dreiländertagung der DGPF, der OVG und der SGPF in Wien, Österreich – Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V., Band **28**, 692-701.

KOVASHKA, A., RUSSAKOVSKY, O., FEI-FEI, L. & GRAUMAN, K., 2016: Crowdsourcing in Computer Vision. Found. Trends Comput. Graph. Vis. **10**(3), 177-243.

NIEMEYER, J., ROTTENSTEINER, F. & SOERGEL, U., 2014: Contextual classification of LiDAR data and building object detection in urban areas. ISPRS Jour. Photogramm. Remote Sens., 87, 152-165.

SCHMOHL, S. & SÖRGEL, U., 2019: Submanifold Sparse Convolutional Networks for Semantic Segmentation of Large-Scale ALS Point Clouds. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. **IV**-2/W5, 77-84.

SETTLES, B., 2009: Active Learning Literature Survey. Computer Sciences Technical Report, 1648, University of Wisconsin-Madison.

SHANNON, C., 1948: A Mathematical Theory of Communication. The Bell System Technical Journal **27**(3), 379-423.

WALTER, V. & SÖRGEL, U., 2018: Implementation, Results, and Problems of Paid Crowd-Based Geospatial Data Collection. Journal of Photogrammetry, Remote Sensing and Geoinformation Science **86**(3-4), 187-197.

WEBLABCENTER INC., 2019: Microworkers. URL: https://www.microworkers.com (letzter Zugriff 20.10.2019).

WORTMAN VAUGHAN, J., 2018: Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. Mach. Learn. Res. **18**(193), 1-46.