

Westfälische Wilhelms-Universität Münster

Institut für Informatik

Praktikum *Computer Vision: Camera Trap Challenge* – Wintersemester 2018/19

Dozenten: Jun.-Prof. Dr. Benjamin Risse, Andreas Nienkötter

Auswertung von Kamerafallenbildern mithilfe von Principal Components Analysis, Spatial Pyramid Matching und Support Vector Machines

Thomas Poschadel
Rudolf-Harbig-Weg 36, 48149 Münster
M.Sc. Informatik
Matrikelnummer: 123 456
blabla@wwu.de

Joschka Strüber
Rudolf-Harbig-Weg 36, 48149 Münster
M.Sc. Informatik
Matrikelnummer: 418 702
j.st@wwu.de

Sufian Zaabalawi
Straße Hausnummer, 12345 Münster
M.Sc. Informatik
Matrikelnummer: 123 456
blabla@wwu.de

Münster, 28. März 2019

Inhaltsverzeichnis

1 Einleitung	1
2 Aufteilung auf Sequenzen	2
2.1 Sortierung mithilfe der Exif-Daten	2
2.2 Camera Trap Sequencer	3
3 Lokalisierung mit Principal Components Analysis (PCA)	4
3.1 Hintergrundapproximation mit PCA	4
3.2 Sliding-Window Lokalisierung mit PCA	6
3.2.1 Objektdetektion mit PCA	6
4 Klassifizierung mit Histograms of Gradients und Support Vector Machines	8
5 Klassifizierung mit Spatial Pyramid Matching	8
5.1 Grundidee Spatial Pyramid Matching	8
5.2 Locality-constrained Linear Coding	9
5.2.1 Grundidee	9
5.2.2 Optimierungsproblem und analytische Lösung	10
5.2.3 Pooling und Normalisierung	11
5.2.4 Klassifizierung mit Support Vector Machines	12
5.3 Implementierung	12
5.3.1 Berechnung der Features	12
5.3.2 LLC Encoding	13
5.3.3 Scikit-learn SPM-Transformer und LinearSVC	13
6 Evaluierung	14
6.1 Daten	14
6.2 Laufzeit Spatial Pyramid Matching	14
6.3 Güte der Klassifizierungstechniken	14
7 Fazit	14
Literaturverzeichnis	15
Eigenständigkeitserklärung	16

1 Einleitung

Kamerafallen bieten eine immer wichtiger werdende Möglichkeiten Populationen zu überwachen und erforschen. Forschungsinteressen sind beispielsweise die Veränderung der Biodiversität, der Einfluss des Klimawandels und anderer Einflüsse auf die Lebensräume und die Migrationsmuster von Populationen [Yu et al. 13].

Durch die immer größer werdende Akzeptanz von Kamerafallen, steigende Qualität und sinkende Preise kommt es zu einer exponentiell wachsenden Datenmenge. Diesen Daten manuell Herr zu werden stellt eine Herausforderung dar, denn jedes Bild muss einzeln von einem Experten auf Tiere untersucht werden. Aufgrund der Verwendung von unveröffentlichten Daten in aktuellen Forschungsprojekten ist Crowd-Sourcing oft unmöglich. Zudem zeichnen sich die anfallenden Bilder durch einen hohen Anteil von False-Positives (Bilder ohne Tiere) und eine große Vielfalt von Arten in verschiedensten Posen, Entferungen und bei wechselnden Witterungsbedingungen aus, was die automatische Analyse erschwert.

Im Kontext dieser Arbeit beziehen wir uns dabei auf einen Datensatz aus dem niederländischen Nationalpark De Hoge Veluwe. Die Datenbank umfasst 40 GB Bilder von neun einheimischen Tierspezies, die mithilfe von Reconyx-Kamerafallen gesammelt wurden. Dabei wurden sowohl Farbbilder am Tag als auch Infrarotbilder in schwarz-weiß in der Nacht aufgenommen.

Um diese komplexe Aufgabe zu automatisieren, stellen wir eine Softwarepipeline vor, mit deren Hilfe es möglich ist alle nacheinander anfallenden Herausforderungen zu lösen. Der erste Schritt besteht in der Ordnung der Daten. Dafür haben wir mit dem *Camera Trap Sequencer* eine Software mit grafischer Benutzeroberfläche implementiert, die es dem Benutzer erlaubt nach Tierarten vorsortierte Datenbanken oder einzelne Ordner von Bildern auf zusammenhängende Sequenzen aufzuteilen.

Der nächste Schritt ist die Lokalisierung von Tierarten in Bildern. Das ermöglicht zum einen das Aussortieren von Bildern, die in Wirklichkeit keine Tiere zeigen, und zum anderen die Identifikation von Bildausschnitten, die für die spätere Klassifizierung relevant sind. Hierfür verwenden wir eine Pipeline mit verschiedenen Vor- und Nachbereitungsschritten, die mithilfe von *Principal Components Analysis* ein Hintergrundbild auf einer Bildsequenz berechnet und somit die Segmentierung von relevanten Bildausschnitten erlaubt. Um auch die Lokalisierung von Tieren auf Einzelbildern zu ermöglichen, wurde zusätzlich ein PCA-unterstütztes *Sliding-Window*-Verfahren implementiert.

Den Abschluss jeder Auswertung bildet das Klassifizieren von Spezies in zuvor bestimmten *Regions of Interest*. Hierzu stellen wir zwei verschiedene Techniken vor:

Die erste ist die Klassifizierung mit Hilfe einer *Support Vector Machine* mit *Radial-Basis-Function*-Kernel auf dem *Histogram-of-oriented-Gradients*-Feature, einem Strukturfeature. Das zweite Verfahren ist *Spatial Pyramid Matching* (SPM) mit *Locality-constrained linear Coding* [LSP06]. Hierbei wird das Eingabebild in immer feinere Teilbilder unterteilt, auf denen dann SIFT- oder

LBP-Features berechnet werden. Diese Features werden mit LLC kodiert, wobei die räumliche Aufteilung erhalten bleibt. Abschließend werden die so bestimmten Codes zum Trainieren einer Support Vector Machine mit linearem Kernel benutzt, da sich empirisch erwiesen hat, dass sie gut linear separierbar sind [Yang et al. 09].

Zum Abschluss unserer Ausarbeitung evaluieren wir unsere Verfahren. Betrachtet werden sowohl die Laufzeit der Algorithmen als auch ihre Güte auf den uns zur Verfügung stehenden Daten. Da der ursprüngliche Datensatz zu groß ist, betrachten wir dabei lediglich zwei repräsentative Teilmengen: In der DDD befinden sich Tagbilder von Dachsen und Damhirschen. Die geringe Datenmenge erlaubte schnelle Ergebnisse, ohne grundlegende Probleme, wie beispielsweise unterschiedlich unbalancierte Klassenhäufigkeiten, aus dem Blick zu verlieren. Für die DDD+ haben wir über 2000 Tag- und Nachtbilder von insgesamt sechs verschiedenen Tierarten zusammengestellt. Ziel ist hierbei die Bestimmung der Güte des Verfahrens auf einem komplexen Datensatz.

2 Aufteilung auf Sequenzen

2.1 Sortierung mithilfe der Exif-Daten

Die handelsüblichen Kamerafalle haben einen Sensor, der bei Bewegung auslöst und zunächst zehn Bilder im Abstand von jeweils einer Sekunde schießt. Sollte es am Ende einer Zehnersequenz weiterhin Bewegung im Sichtfeld der Kamera geben, löst der Sensor erneut aus und es werden weitere zehn Bilder aufgenommen. Das sorgt dafür, dass der Datensatz aus zusammenhängenden Sequenzen von jeweils zehn oder mehr Bildern besteht. Oft befinden sich gerade auf den letzten Aufnahmen einer Sequenz keine Tiere mehr, da sich diese aus dem Bild bewegt haben.

Unglücklicherweise sind die Daten auf der Datenbank lediglich nach Tierart sowie dort jeweils nach Tag und Nacht, leeren Bildern und Fehlklassifizierungen sortiert. Für das korrekte Funktionieren unserer Segmentierungstechnik PCA ist es aber nötig, dass die Daten in zusammenhängenden Sequenzen vorliegen. Aus diesem Grund benutzen wir die Exif-Metadaten, um die Daten aufzuteilen. Diese Exif-Daten umfassen Informationen über das Bild, wie beispielsweise die Abmessungen, das Aufnahmedatum, die Belichtungszeit, den ISO-Wert und das Kameramodell.

Unser Algorithmus sammelt zunächst die Metadaten aller aufzuteilenden Bilder in einer Liste. Diese Liste wird primär nach Seriennummer der Kamera und sekundär nach Aufnahmezeitpunkt sortiert. Sollte es zwischen zwei benachbarten Bildern in der Liste zu einem Wechsel der Seriennummer kommen, so wissen wir, dass eine neue Sequenz beginnt. Ebenso gilt das für zwei Bilder, die von derselben Kamera aufgenommen wurden, deren Aufnahmezeitpunkte sich jedoch um mehr als ein paar Sekunden unterscheiden. Diese beiden Situationen markieren den Wechsel einer Sequenz anhand der wir unterscheiden können, welche Bilder zusammenhängen.

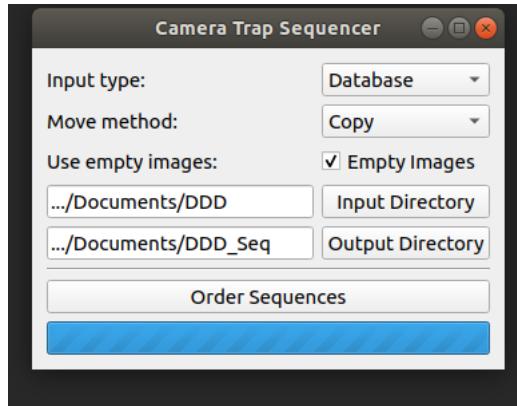


Abbildung 1: Der Camera Trap Sequencer im Linux-Design.

Leider gehört die Seriennummer nicht zu den ursprünglichen Exif-Metadaten. Stattdessen wird sie in die sogenannten *Maker Notes* geschrieben, einem freien Datenfeld, dass die Kamerahersteller für ihre Zwecke benutzen können. Aktuell gibt es keine Python-Bibliothek, die das Maker-Note-Feld auslesen kann. Aus diesem Grund waren wir gezwungen die Perl-Bibliothek „ExifTool“ [Harvey 03] zu benutzen und auf sie mit Hilfe einer Python-Schnittstelle zuzugreifen.

Zugegebenermaßen ist die Verwendung der Metadaten nicht der eleganteste Weg Sequenzen zu bestimmen. Sie hat aber - anders als das Auslesen und Abgleichen der Pixel aus der linken oberen Ecke des Bildes - den Vorteil unabhängig vom Kamerahersteller und -modell zu sein und ist sehr effizient.

2.2 Camera Trap Sequencer

Der Camera Trap Sequencer ermöglicht es dem Anwender seine Datenbank mit Kamerafallenbildern unkompliziert und schnell auf Sequenzen aufzuteilen. Als Eingabe kann sowohl eine komplette Datenbank als auch ein einzelner Ordner mit Bildern benutzt werden. Man kann sich zwischen dem Verschieben und Kopieren der Bilder entscheiden. Das Verschieben von Bildern hat den Vorteil, dass keine Daten dupliziert werden. Die ursprüngliche Ordnerstruktur bleibt momentan jedoch noch erhalten.

Abschließend hat man im Fall einer Bilddatenbank die Möglichkeit sich dafür zu entscheiden mit Informationen über leere Bilder zu speichern. Alle vorsortierten Datenbanken haben Verzeichnisse mit dem Namen „empty“, in denen sich Bilder auf Sequenzen befinden, auf denen keine Tiere zu sehen sind. Unsere Segmentierungstechnik PCA ist darauf angewiesen möglichst auf möglichst langen Bildsequenzen angewendet zu werden. Deshalb ist es nützlich, False-Positives mitzuverwenden, insbesondere weil leere Bilder einen großen Beitrag zur Berechnung eines leeren Hintergrundbilds leisten können. Da wir als Label für die Klassifizierung aber die Namen der Tierordner benutzen, müssen diese vor der Klassifizierung aussortiert werden. Deshalb speichern wir, falls die empty-Option gesetzt ist, eine Textdatei mit den Dateipfaden aller leeren Bilder, die dann im Anschluss an PCA aussortiert werden können.

Der Camera Trap Sequencer selbst ist mit PyQt5 umgesetzt. Er zeichnet sich deshalb durch ein natives Design auf jeder Plattform aus.

3 Lokalisierung mit Principal Components Analysis (PCA)

Die automatische Lokalisierung von Objekten in digitalen Bildern ist ein wesentlicher Bestandteil vieler Anwendungen. Für das Lokalisierungsproblem in dieser Arbeit bietet sich die Verwendung der Methoden *Hintergrund-Subtraktion* und *Sliding-Window* mit PCA an.

3.1 Hintergrundapproximation mit PCA

Um Bewegungen in Bildsequenzen erkennen zu können, wird in der Praxis sehr häufig das Verfahren der *Hintergrund-Subtraktion* angewandt. Dabei handelt es sich um ein klassisches Verfahren aus dem Bereich der Bilderkennung. Das Hintergrundbild kann mithilfe von PCA approximiert werden. Anschließend wird das Vordergrundbild über die Differenz zum Hintergrundbild extrahiert. PCA, oder auch Hauptkomponentenanalyse, ist ein statistisches Verfahren um große Mengen von Datensätzen zu vereinfachen und zu strukturieren, indem die Datenpunkte im p -dimensionalen Raum R^p in einen q -dimensionalen Unterraum R^q mit ($q < p$) projiziert werden. Diese Transformation muss dabei so gewählt werden, dass möglichst wenig Information verloren geht. Grundsätzlich benutzt PCA die *Niedrigrangapproximation*. Damit kann eine Matrix durch eine andere Matrix im allgemeinen Rang angenähert werden. Sei eine Matrix A mit $\text{Rang}(A) = r$ und $r > k$:

$$\min_{\text{rang}(A)=k} \|A - B\|_2 \quad (1)$$

Dabei soll die Differenz zwischen A und B minimiert werden. Mithilfe der *Singulärwertzerlegung* (SVD) können die Singulärwerte einer Matrix abgelesen werden. Die SVD von Matrix A ist dann:

$$A = U\Sigma V^T \quad (2)$$

Somit kann ein Hintergrundbild aus einer Sequenz von Bildern wie folgt approximiert werden (Abbildung 2):

- ▷ Berechne Singulärwertzerlegung aller Bildern von Sequenz X :

$$SVD(X) = C = U\Sigma V^T \quad (3)$$

- ▷ Leite die Matrix Σ_k von Σ her, sodass die Werte $n - k$ entlang der Diagonale durch 0 ersetzt werden.

- ▷ Dies ergibt die Niedrigrangapproximation von Matrix X :

$$SVD(X)_k = C_k = U\Sigma_k V^T \quad \text{mit} \quad \Sigma_k = \text{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0) \quad (4)$$

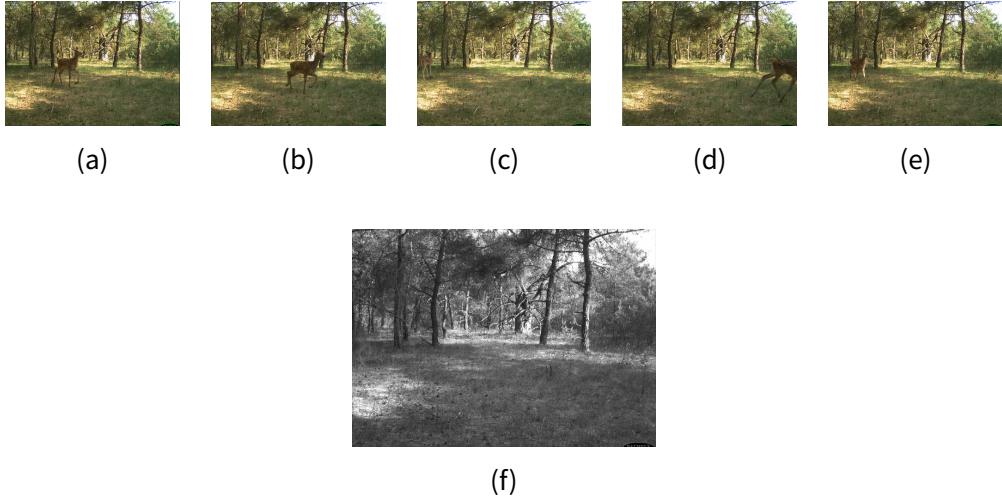


Abbildung 2: (a)-(e) Bilder aus einer Sequenz und (f) das approximierte Hintergrundbild.

Anschließend kann das Vordergrundbild durch die klassische *Hintergrund-Subtraktion* extrahiert werden (Abbildung 3).

The diagram illustrates the foreground extraction process. It shows a grayscale image of a deer in a forest (M) minus a grayscale background approximation (L) equals the foreground image (S). Below the equation, the letters M, -, L, =, and S are written in a large, bold font.

Abbildung 3: Das Vordergrundbild S ergibt sich durch die Subtraktion des approximierten Hintergrundbildes L .

Zum Nachbearbeitung des Vordergrundes gehört eine Vielzahl von Operationen z.B. *morphologische Operationen*, *Thresholding* und *Filterung*. Damit können kleinere Bildstrukturen und Rauschen entfernt, vergrößert, geschlossen oder aufgefüllt werden. Können jedoch diese Operationen zu einer Veränderung der Größe der Vordergrundelemente führen, was zur Lokalisierung des Elements aber keinen Störfaktor ergibt. Durch Kombination der Operationen in einer bestimmten Reihenfolge kann Größenveränderung verhindert und dennoch die Vorteile der Operationen genutzt werden. Durch *Opening* werden zunächst kleine Strukturen bzw. Rauschen, welches zum Hintergrund gehört, entfernt. Danach werden kleine Löcher innerhalb der Vordergrundelemente durch *Closing* geschlossen. In (Abbildung 4) ist eine Kombination dieser Pipeline zur Nachbearbeitung des Vordergrundes benutzt worden.

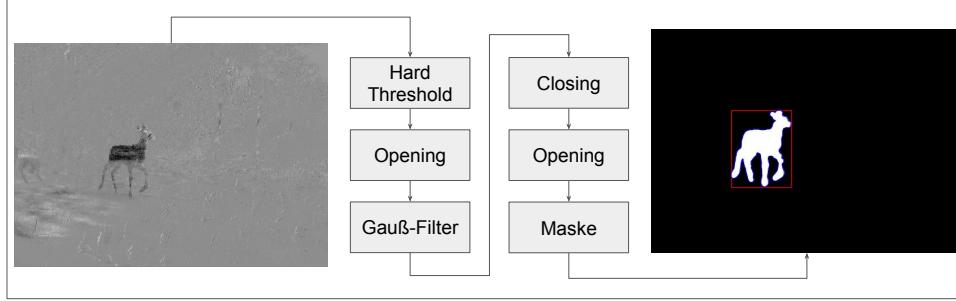


Abbildung 4: Die Pipeline der Nachbearbeitung des Vordergrundbildes. Durch *Opening* und *Closing* werden kleine Bildstrukturen bzw. Rauschen entfernt und kleine Löcher geschlossen werden. Die Gauß-Filterung dient in diesem Fall dazu, die Silhouette des Vordergrundelements grob zu vergrößern.

3.2 Sliding-Window Lokalisierung mit PCA

Sliding-Window ist eine Brute-Force-Suche über das Bild mit fester Fenstergröße, um Objekte zu finden. Für jedes dieser Fenster wird ein Bildklassifikator angewendet, um zu bestimmen, ob das Fenster ein bekanntes Objekt enthält. In diesem Fall wird PCA als Objekt-Klassifikator angewandt.

3.2.1 Objektdetektion mit PCA

Jedes Bild ist ein Punkt in einem hochdimensionalen Raum. Durch das PCA-Verfahren lassen sich die Datenpunkte in einen niederdimensionalen Unterraum abbilden. PCA sucht die ersten k -Hauptkomponenten, welche die Daten mit einer maximalen Varianz beschreiben. Damit wird es eine niederdimensionale Darstellung gefunden, bei der die Klassifizierung leichter wird.

Algorithmus

- ▷ Phase I: Initialisierung
 - ▷ Berechne das Mittelwertbild der Trainingsbilder

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (5)$$

- ▷ Berechne die zentrierten Daten durch Subtraktion der Trainingsbilder vom Mittelwertbild

$$C = X - \mu \quad (6)$$

- ▷ Berechne die Eigenwerte und Eigenvektoren für die Kovarianzmatrix CC^T

$$\text{SVD}(C) = \mathbf{U}\Sigma\mathbf{V}^T \quad (7)$$

▷ Projiziere die Trainingsbilder in den r -Unterraum

$$\mathbf{Y} = \mathbf{U}_r^T C \quad (8)$$

▷ Phase II: Klassifikation

Gegeben ist ein unbekanntes Bild M

▷ Projiziere das Bild M in den r -Unterraum

$$\mathbf{W} = \mathbf{U}_r^T (M - \mu) \quad (9)$$

▷ Finde den nächsten Nachbarn zwischen den projizierten Trainingsbildern \mathbf{Y} und dem projizierten Bild \mathbf{W} .

Die Sliding-Windows laufen das Bild mit unterschiedlichen Fenstergrößen durch. Demnach werden Schnittbilder einzelne mit PCA klassifiziert. Dabei wird der nächste Nachbar der projizierten Schnittbilder gefunden und zugeordnet (Abbildung 5).



Abbildung 5: Lokalisierung mit Sliding-windows und PCA.

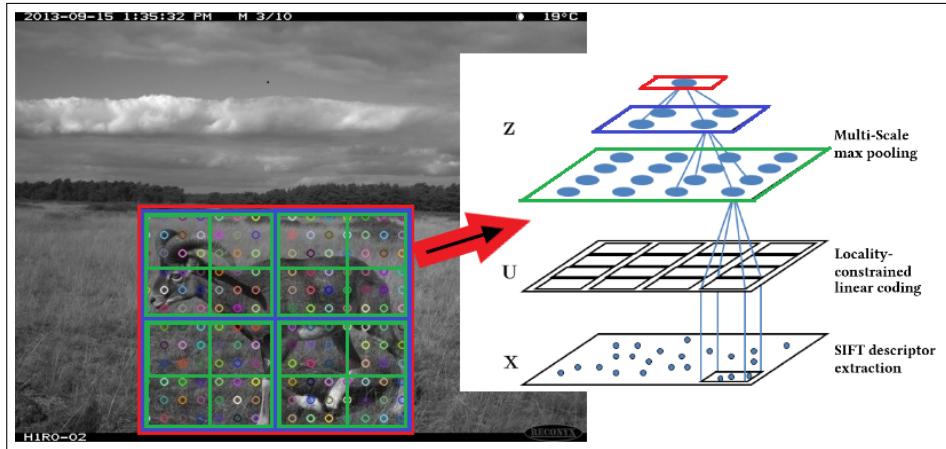


Abbildung 6: Architektur des Algorithmus mit Veranschaulichung der Spatial Pyramid auf SIFT-Features und Pooling. Im Anschluss werden der L_0 - (rot), die vier L_1 - (blau) und 16 L_2 -Codes (grün) konkateniert - angelehnt an [Yang et al. 09].

4 Klassifizierung mit Histograms of Gradients und Support Vector Machines

5 Klassifizierung mit Spatial Pyramid Matching

5.1 Grundidee Spatial Pyramid Matching

Spatial Pyramid Matching ist eine Weiterentwicklung des Bag-of-visual-Words-Ansatzes von Lazebnik et al. [LSP06]. Er wurde 2006 eingeführt und ursprünglich zur Klassifizierung von Szenerien benutzt, um zu erkennen, ob ein Bild beispielsweise eine Stadt, einen Wald oder einen Strand zeigt. Beim Spatial Pyramid Matching werden nicht einfach alle Features ungeordnet betrachtet. Stattdessen wird das Bild in immer feinere Teilbilder unterteilt, deren Features jeweils in einem Spatial Bin gesammelt werden. Der L_0 -Bin wird in jeder Dimension halbiert und bildet somit vier L_1 -Bins. Diese wiederum werden wieder geviertelt, was insgesamt 16 L_2 -Bins bedeutet.

Die Features aus jedem der insgesamt 21 Bins werden anschließend über die Lösung eines Optimierungsproblems den Features eines Codebooks B zugeordnet. Bei dem Codebook handelt es sich um eine Menge von Features, die den Featureraum über der Datenmenge gut widerspiegelt. Dazu werden üblicherweise Features über zufälligen Bildern oder Bildausschnitten der Datenbank berechnet und anschließend mithilfe von Clustering Repräsentanten bestimmt. Bei einigen Verfahren wird das Codebook auch online weiter optimiert, worauf wir jedoch verzichten. Wir haben in unseren Evaluierungen Codebooks mit 256, 512, 1024 und 2048 Features getestet. Je größer das Codebook ist, desto besser können auch komplizierte Datenbanken mit vielen und vielfältigen Klassen abgebildet werden. Auf der anderen Seite steigt die Komplexität der Berechnungen mit größerem Codebook deutlich an.

Nachdem für jeden Spatial Bin alle Features mit dem Codebook kodiert wurden, werden diese pro Bin gepoolt, um einen einzigen Zuordnung für diesen zu bilden. Die Länge entspricht dabei der Anzahl der Features im Codebook. Abschließend werden die Kodierungen aller Bins konkateniert. Für ein Codebook mit 1024 Features ergibt sich so beispielsweise ein SPM-Code der Länge $21 \cdot 1024 = 21504$. Diese SPM-Codes können dann mit Klassifizierern wie beispielsweise Support Vector Machines (SVMs) eingesetzt werden. Durch die Länge der Codes könnte man erwarten, dass die Laufzeit der Klassifizierung sehr hoch ist. Wie wir im nächsten Abschnitt sehen werden, sind diese gut linear separierbar, weshalb eine SVM mit effizientem linearen Kernel verwendet werden kann. Ein Überblick über das Verfahren befindet sich in Abbildung 6.

Unser Algorithmus orientiert sich insgesamt lose an dem Verfahren von [Yu et al. 13]. Dort wurden zunächst SIFT- und cLBP-Features dicht auf Bildern berechnet. Mit dem Feature Sign Solver haben sie dann ein dünnbesetztes Kodierungsproblem gelöst [Lee et al. 07]. Die Codes wurden mit Max-Pooling gepoolt, mit der euklidischen Norm normalisiert und anschließend wurden die Bins konkateniert. Die unabhängig voneinander kodierten SIFT und cLBP-Features wurden abschließend mithilfe von AdaBoost und einer linearen SVM klassifiziert.

5.2 Locality-constrained Linear Coding

5.2.1 Grundidee

Wir verwenden zur Kodierung unserer Features das *Locality-constrained Linear Coding* (LLC) [Wang et al. 10]. Bei diesem werden jedem Feature f der Spatial Bins mehrere Features des Codebooks zugeordnet, die insgesamt keinen zu großen euklidischen Abstand von f haben dürfen. Bei der ursprünglichen Variante des Spatial Pyramid Matchings wurde lediglich Vektorquantisierung benutzt [LSP06]. Das heißt, dass jedes f dem nächsten Nachbarn im Codebook zugewiesen wird. Unglücklicherweise kommt es dabei zu schlechten Zuordnungen, wenn es beispielsweise keine Feature im Codebook gibt, das zu f ähnlich ist (1) oder wenn zwei recht ähnliche Features f und g unterschiedlichen Features in B zugeordnet werden (2). Fehler dieser Art nennt man *Quantisierungsfehler*.

Eine deutliche Verbesserung stellte das Verfahren von [Yang et al. 09]. Beim *SPM based on sparse coding* (ScSPM) wird jedes Feature f anteilig gleich mehreren Features in B zugeordnet. Über einen Regularisierungsterm wird dabei sichergestellt, dass die Kodierungen insgesamt dünnbesetzt sind. Dieser Algorithmus bietet eine deutliche Verbesserung bei Fehlertyp (1), denn möglicherweise lässt sich f durch Kombination mehrerer Features in B besser darstellen. Da lediglich die Dünnsbesetztheit (*Sparsity*) der Kodierungen gefordert ist, können weiterhin zwei ähnliche Features durch vollkommen unterschiedliche Kombinationen von Features aus B repräsentiert werden.

Wang et al. haben festgestellt, dass für eine optimale Zuordnung zum Codebook Sparsity allein nicht ausreicht [Wang et al. 10]. Deshalb stellen sie in ihrem Optimierungsproblem über einen

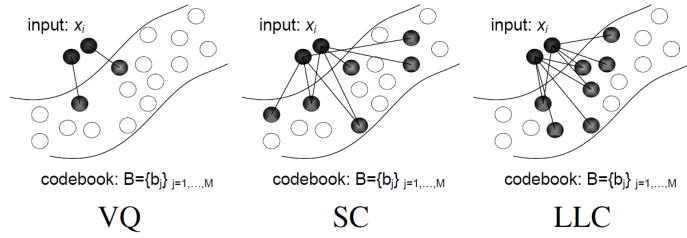


Abbildung 7: Vergleich der drei Kodierungsstrategien Vector Quantization, Sparse Coding und Locality-constrained Linear Coding [Wang et al. 10].

Regularisierungsterm zusätzlich sicher, dass die Zuordnung lokal stattfindet. Die Lokalität stellt gleichzeitig auch die Sparsity sicher, während das Gegenteil nicht immer der Fall ist. Ein Vergleich dieser Strategien befindet sich in Abbildung 7.

5.2.2 Optimierungsproblem und analytische Lösung

Die Beschreibung von LLC folgt der Veröffentlichung von [Wang et al. 10]. Es seien eine Menge von D-dimensionalen Features $X = [x_1, \dots, x_N] \in \mathbb{R}^{(D \times N)}$ und ein Codebook $B = [b_1, \dots, b_M] \in \mathbb{R}^{(D \times M)}$ mit M Einträgen gegeben. Bei X handelt es sich um Features aus einem Bild und bei B um eine Menge von Deskriptoren, die den Merkmalsraum über allen Bildern gut widerspiegelt.

Gesucht ist eine lokale Zuordnung $C = [c_1, \dots, c_N] \in \mathbb{R}^{(M \times N)}$ von Features aus X auf Visual Words aus B . Hierfür verwenden wir das folgende Optimierungsproblem:

$$\min_C \sum_{i=1}^N \|x_i - B \cdot c_i\|^2 + \alpha \cdot \|d_i \odot c_i\|^2, \text{ s.t. } 1^T c_i = 1 \forall i \quad (10)$$

Betrachten wir den Regularisierungsterm $\alpha \cdot \|d_i \odot c_i\|^2$ genauer. \odot ist die elementweise Multiplikation zweier Vektoren. $d_i = \exp\left(\frac{\text{dist}(x_i, B)}{\sigma}\right) \in \mathbb{R}^M$ hingegen ist ein Vektor, der die Distanz des Features x_i zu allen Visual Words aus B angibt. Dabei ist $\text{dist}(x_i, B) = [l_2(x_i, b_1), \dots, l_2(x_i, b_m)]^T \in \mathbb{R}^M$ der Vektor der euklidischen Distanzen von x_i zu jedem Visual Word aus B . In der Praxis wird d_i zusätzlich normalisiert, indem das Maximum aller $l_2(x_i, b_j)$ bestimmt wird. Dieses wird von jeder Zeile von $\text{dist}(x_i, B)$ abgezogen, wodurch sich Werte im Intervall $(-\infty, 0]$ ergeben. Die Anwendung der Exponentialfunktion sorgt dann für normalisierte Distanzwerte in $(0, 1]$. α und σ sind Hyperparameter, die bestimmen wie lokal die Zuordnungen sein müssen.

Es ist zu beachten, dass die Kodierungen c_i nicht zwangsläufig im Sinne der l_0 -Norm dünnbesetzt sind. Vielmehr ergeben sich nicht viele relevante Werte, sodass alle zu kleinen Koeffizienten mithilfe eines Thresholds auf 0 gesetzt werden können, um echte Sparsity sicherzustellen.

Anders als Sparse Coding besitzt LLC eine analytische Lösung und muss nicht mit dem Feature Sign Solver gelöst werden. Als erstes wird die Kovarianzmatrix eines Features x_i über dem

Codebook B berechnet, indem wir x_i zeilenweise von B abziehen und die resultierende Matrix mit dem Transponierten seiner selbst multiplizieren:

$$C_i = (B - 1 \cdot x_i^T) \cdot (B - 1 \cdot x_i^T)^T \quad (11)$$

Die Kovarianzmatrix C_i benutzen wir, um ein lineares Gleichungssystem über dieser Matrix und dem Regularisierungsterm nach \tilde{c}_i zu lösen und anschließend zu normalisieren:

$$(C_i + \lambda \operatorname{diag}(d_i)) \cdot \tilde{c}_i = (1, \dots, 1)^T \quad (12)$$

$$c_i = \tilde{c}_i / \tilde{c}_i^T \tilde{c}_i \quad (13)$$

[Wang et al. 10] geben die analytische Lösung von LLC als Vorteil gegenüber Verfahren an, die den Feature Sign Algorithmus benutzen, da dieser im besten Fall ein Laufzeit in $\mathcal{O}(M \cdot K)$ hat, wobei K die Anzahl der Elemente ungleich Null ist. Da C_i nicht dünnbesetzt ist, hat die Lösung des linearen Gleichungssystems im Allgemeinen eine Komplexität von $\mathcal{O}(M^3)$. Möglicherweise lässt sich das Gleichungssystem mit einem speziellen Solver schneller lösen, da die Kovarianzmatrix symmetrisch und positiv semidefinit ist. In der Praxis hat sich dieser Schritt aber als Flaschenhals des Verfahrens herausgestellt, wie wir in Abschnitt 6 feststellen werden.

5.2.3 Pooling und Normalisierung

Das Resultat von LLC auf eine Menge Featurs $X \in \mathbb{R}^{(D \times N)}$ ist eine Menge von Kodierungen $C \in \mathbb{R}^{N \times M}$. C wird auf einen einzelnen trainierten Deskriptor abgebildet, indem wir sie spaltenweise poolen:

- ▷ sum pooling: $c_{out} = \sum_{i=1}^N c_{in,i}$
- ▷ max pooling: $c_{out} = \max c_{in,1}, \dots, c_{in,N}$

Das Pooling stellt dabei Möglichkeit dar, die Informationen aller Kodierungen aus C in einem Deskriptor zu bündeln und die aussagekräftigsten Informationen dabei zu erhalten. Aufgrund der Assoziativität der Pooling-Operationen ist es nicht nötig alle 21 Bins einzeln zu kodieren. Das erspart einen Menge Aufwand, da jedes Feature aus einem L_2 -Bin auch in einem L_1 - und dem L_0 -Bin auftaucht. Stattdessen reicht es alle L_2 -Bins einzeln zu kodieren, zu poolen und danach sukzessive das Pooling auf die vier jeweils zusammengehörigen Bins anzuwenden. Dieses *Multiscale Pooling* wurde auch in Abbildung 6 veranschaulicht.

Zum Abschluss des LLC-Verfahrens werden alle 21 Deskriptoren der einzelnen Bins zu einem einzigen langen LLC-Code konkateniert. Um die Vergleichbarkeit zwischen diesen zu gewährleisten, normalisieren wir sie mit einer der folgenden Methoden:

- ▷ sum normalization: $c_{out} = c_{in} / \sum_j c_{in}(j)$
- ▷ l^2 normalization: $c_{out} = c_{in} / \|c_{in}\|_2$

5.2.4 Klassifizierung mit Support Vector Machines

Die LLC-Codes lassen sich mit einer Support Vector Machine (SVM) mit linearem Kernel klassifizieren. Diese separiert die Klassen mit einer linearen Hyperebene. Ohne auf die Details einzugehen verwenden wir als Loss-Funktion den differenzierbaren *Squared Hinge Loss* und im Falle von mehr als zwei Klassen werden mit der *One-against-all*-Strategie L verschiedene Klassifikatoren trainiert.

Der von uns verwendete lineare Kernel der SVM hat den Vorteil, dass wir den Klassifikator in $\mathcal{O}(N)$ Zeit trainieren können, während das Testen einen Codes sogar in $\mathcal{O}(1)$ Zeit möglich ist [Yang et al. 09]. Die Laufzeit wird lediglich von der Dimensionalität der Daten, also der Größe des Codebooks bestimmt, nicht aber von der Gesamtanzahl der Features. Demgegenüber benötigen nichtlineare Mercer-Kernels wie beispielsweise der *Chi-square Kernel* oder der im vorigen Kapitel verwendete RBF-Kernel Laufzeiten von $\mathcal{O}(N^2)$ bis $\mathcal{O}(N^3)$ fürs Training beziehungsweise $\mathcal{O}(N)$ fürs Testen. Diese Erkenntnis deckt sich auch mit unseren Evaluierungen, in der das Training der SVM für die Laufzeit des Verfahrens nicht relevant ist. Diese wird stattdessen von der Laufzeit des LLC dominiert.

5.3 Implementierung

Die Implementierung des Verfahrens erfolgte in der Programmiersprache Python, wobei verschiedene Bibliotheken zum Einsatz gekommen sind. Zu diesem Zeitpunkt seien uns die Pfade der Bilder und die Regions of Interest der zu klassifizierenden Tiere als Bounding Boxes gegeben. Um die Laufzeit des Verfahrens zu verringern, werden alle Teilbilder auf eine maximale Länge oder Breite von 300 Pixeln verkleinert und in Graustufenbilder umgerechnet. Das Seitenverhältnis bleibt konstant.

5.3.1 Berechnung der Features

Wir verwenden zur Klassifizierung der Bilder zwei verschiedene Arten von Features. Das *Scale-invariant-Feature-Transform*-Feature (SIFT) ist ein Strukturfeature, bei dem über eine *Scale-space*-Pyramide von *Difference-of-Gaussian*-Bildern Histogramme von Gradienten berechnet werden [Lowe 99]. Wir benutzen die Implementierung von OpenCV, die sich in dem Modul `opencv-contrib` der patentierten Verfahren befindet und gegebenenfalls nicht von sich aus mit OpenCV installiert wird [**ocv**]. SIFT findet vielfältig Verwendung und zeichnet sich durch hohe Aussagekraft sowie die Invarianz gegenüber Skalierungen und geringen Perspektiv- und Belichtungswechseln aus.

Das zweite Feature ist das Texturfeature *Local Binary Binary* (LBP) [OPH 94]. Wir verwenden die uniforme Variante mit Radius zwei und 16 benachbarten Punkten von „Scikit-image“ [SK]. Hierbei wird ein Graustufenpixel mit 16 symmetrisch verteilten Pixel im Abstand von zwei Pixeln verglichen. Falls der Nachbar größer ist als der zentrale Pixel, merken wir uns für diese Stelle eine 0, ist er kleiner, eine 1. Das ergibt bei 16 Nachbarn eine Kodierung von zwei Bytes. In der uniformen Variante prüfen wir zusätzlich, ob der Deskriptor uniform ist, also ob es höchstens zwei Transformationen der Form $0 \rightarrow 1$ oder $1 \rightarrow 0$ gibt. Bei der anschließenden Berechnung des Histogramms über einem Block von Pixeln erstellen wir für jedes uniforme Muster einen Bin im Histogramm und einen einzelnen Bin für alle nicht-uniformen Muster. Die Verwendung der uniformen Variante sorgt dafür, dass die Histogramme lediglich die Länge 18 statt 512 haben und sorgen zusätzlich für Graustufen- und Rotationsinvarianz.

Die Klasse `FeatureExtraction` ermöglicht das Berechnen von Featuremengen und Features in Spatial Pyramids für übergebene Featureextraktoren wie beispielsweise dem für SIFT oder LBP. Die Features werden in einem dichten Gitter der Schrittweite 16 über Blöcken von 16 mal 16 Pixeln berechnet. Für die Klassifizierung wäre eine geringere Schrittweite (beispielsweise vier Pixel) vermutlich besser, da bei nicht überlappenden Blöcken wichtige Kanten und Eckpunkte verloren gehen können. Dieses Verfahren wurde auch von [Yu et al. 13] verwendet. Wir verzichten bewusst darauf, um die Anzahl der Features und damit die Komplexität der Berechnungen zu verringern.

5.3.2 LLC Encoding

Den Kern unseres Verfahrens bildet die Klasse `LlcSpatialPyramidEncoder`, der über die `encode`-Methode Features mit LLC kodieren kann. Zuerst wird mithilfe von Scikit-learns `MiniBatchKMeans` ein Codebook über einer Menge von Features trainiert, die auf der Datenbank der Bilder zufällig bestimmt wurden [SKL]. Anschließend wurde der Algorithmus aus Abschnitt 5.3.3 mit „Numpy“ umgesetzt. Leider hat sich die Laufzeit der Kodierung als problematisch herausgestellt (s. 6), weshalb der ursprünglich sequentielle Algorithmus in klassischem Python mehrmals optimiert wurde. Die Hilfsfunktionen der Klasse wurden in das Modul `llc_optimization` ausgelagert und mit „Numba“ annotiert [Numba]. Numba bietet die Möglichkeit Pythonfunktionen, die lediglich die Standardbibliothek und Numpy verwenden, mit sogenannten *Decorators* zu markieren. Bei der Ausführung des Programms werden diese Funktionen mit Hilfe des LLVM-Compilers in Echtzeit kompiliert, was in einer deutlich schnelleren Laufzeit im Vergleich zu interpretiertem Code resultiert.

5.3.3 Scikit-learn SPM-Transformer und LinearSVC

Alle obigen Funktionen wurden in der Klasse `SpmTransformer` zusammengefasst. Diese implementiert als Unterklasse von Scikit-learns `BaseEstimator` die Methoden `fit` und `transform`, wodurch sich das LLC-Verfahren problemlos in Scikit-learns *Pipelines* integrieren lässt [SKL]. Das

ermöglicht einem spielend leicht die Kombination mit Klassifizierern, Ensemblemethoden und Modelselektionsverfahren der Bibliothek.

Abschließend wurde in dem Modul `spatial_pyramid_matching` das komplette Verfahren von der Segmentierung, über Locality-constrained Linear Coding bis zur Klassifizierung mit Scikit-learns `LinearSVC` auf zwei verschiedenen Datenbanken umgesetzt. `LinearSVC` setzt glücklicherweise bereits alle unsere Anforderungen bezüglich des Kernels, der Loss-Funktion und der Strategie für nicht-binäre Klassifikation um. Um der unbalancierten Datenlage Genüge zu leisten, setzen wir den Parameter `class_weight='balanced'`. Das sorgt dafür, dass der Strafterm `C` invers proportional zum Anteil der Label jeder Klasse modifiziert wird.

6 Evaluierung

-hinweis auf laufzeit und klassifizierung -pca nicht testbar, da keine ground-truth bilder

6.1 Daten

-hoge veluwe -kamerafallenbilder, sequenzen -größe -spezies -DDD/DDD+

6.2 Laufzeit Spatial Pyramid Matching

-sequentiell -multiprocessing -numba -hauptproblem: lösung des gleichungssystems -ausblick: umsetzung mit tensorflow

6.3 Güte der Klassifizierungstechniken

-HOG: auf DDD und DDD+ -SPM: -random search cv mit fünf folds -scipy reciprocal -versch. Größen -SIFT Güte auf DDD -LBP Güte auf DDD -SIFT Güte auf DDD+ -kombination von SIFT und LBP auf DDD+

7 Fazit

-kurzbeschreibung der verfahren und ergebnisse -ausblick: -tensorflow: schnellere laufzeit -> training mit größeren codebooks auf mehr daten -ensemblemethoden zur besseren kombination von sift und clbp: soft voting (SVC, statt LinearSVC), boosting

Literatur

- [Harvey 03] *ExifTool by Phil Harvey.* <https://www.sno.phy.queensu.ca/~phil/exiftool/>. Accessed: 04.03.2019.
- [Lee et al. 07] Honglak Lee, Alexis Battle u. a. „Efficient sparse coding algorithms“. In: *NIPS Proceedings 2007*. NIPS, 2007.
- [Lowe 99] David Lowe. „Object recognition from local scale-invariant features“. In: *Proceedings of the International Conference on Computer Vision*. IEEE, 1999.
- [LSP06] Svetlana Lazebnik, Cordelia Schmid u. a. „Beyond bags of features: spatial pyramid matching for recognizing natural scene categories“. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2006.
- [Numba] *Numba.* <http://numba.pydata.org/>. Accessed: 04.03.2019.
- [OPH 94] Timo Ojala, Mati Pietikäinen u. a. „Performance evaluation of texture measures with classification based on Kullback discrimination of distributions“. In: *Proceedings of the 12th IAPR International Conference on Pattern Recognition*. Elsevier, 1994.
- [SKI] *Scikit-learn. scikit-image.org*. Accessed: 28.03.2019.
- [SKL] *Scikit-learn. scikit-learn.org*. Accessed: 04.03.2019.
- [Wang et al. 10] Jinjun Wang, Jianchao Yang u. a. „Locality-constrained Linear Coding for Image Classification“. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.
- [Yang et al. 09] Jianchao Yang, Kai Yu u. a. „Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification“. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.
- [Yu et al. 13] Xiaoyuan Yu, Jiangping Wang u. a. „Automated identification of animal species in camera trap images“. In: *EURASIP Journal on Image and Video Processing* (2013).

Eigenständigkeitserklärung

Hiermit versichern wir, dass die vorliegende Ausarbeitung *Auswertung von Kamerafallenbildern mit Hilfe von Principal Components Analysis, Spatial Pyramid Matching und Support Vector Machines* selbstständig verfasst worden ist, dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken – auch elektronischen Medien – dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind.

(Ort, Datum)

(Unterschrift)