Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Cybernetics
Study program: Open Informatics

Rekonstrukce trojdimenzionálních modelů zvířat

Reconstructing 3D Models of Animals

MASTER THESIS

Author: Bc. Valeriia Iegorova
Supervisor: Doc. Ing. Tomáš Pajdla, Ph.D.
Submitted: January 2022

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Iegorova Valeriia**	Personal ID number:	**453535**
Faculty / Institute:	**Faculty of Electrical Engineering**		
Department / Institute:	**Department of Cybernetics**		
Study program:	**Open Informatics**		
Specialisation:	**Computer Vision and Image Processing**		

II. Master’s thesis details

Master’s thesis title in English:

Reconstructing 3D Models of Animals

Master’s thesis title in Czech:

Rekonstrukce trojdimenzionálních modelů zvířat

Guidelines:

1) Study the approaches to 3D reconstruction from images and 3D data relevant for reconstructing non-static animals [1,2,3,4,5].
2) Experiment with available 3D reconstruction pipelines and evaluate their suitability for your task [1].
3) Choose the most promising approach and adjust it to you task.
4) Demonstrate the results of the adjusted approach on real data.

Bibliography / sources:

[1] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, Michael J. Black – 3D Menagerie: Modeling the 3D Shape and Pose of Animals - 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
[2] Richard A. Newcombe, D. Fox, S.Seitz - DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time - 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
[3] Silvia Zuffi, Angjoo Kanazawa, T. Berger-Wolf, Michael J. Black - Three-D Safari: Learning to Estimate Zebra Pose, Shape, and Texture From Images “In the Wild” - 2019 IEEE/CVF International Conference on Computer Vision (ICCV)
[4] Aljaž Božič, Michael Zollhöfer, Christian Theobalt, Matthias Nießner - DeepDeform: Learning Non-rigid RGB-D Reconstruction with Semi-supervised Data - 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
[5] Richard Hartley, Andrew Zisserman – Multiple View Geometry in Computer Vision – 2003 Cambirdge Univerersity Press

Name and workplace of master’s thesis supervisor:

doc. Ing. Tomáš Pajdla, Ph.D., Applied Algebra and Geometry, CIIRC

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **12.02.2021** Deadline for master's thesis submission: **04.01.2022**

Assignment valid until: **30.09.2022**

doc. Ing. Tomáš Pajdla, Ph.D. Supervisor’s signature	prof. Ing. Tomáš Svoboda, Ph.D. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

_______ . _______________ Date of assignment receipt

______________________________ Student’s signature

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, January 2022

Bc. Valeriia Iegorova

Acknowledgements

I would like to thank my friends and colleagues Mgr. Juraj Páll and Ing. Jakub Jirůtka for their support and for review of this work. I would like to express special thanks to Ing. Tomáš Borovička for giving me access to computational resources and supporting this research, and to MSc. Jan Bím PhD for his valuable comments. Also, I thank Dr. Silvia Zuffi, the author of the *SMAL* animal model, and Benjamin Biggs, the author of the *SMBDL* model, for discussing their research with me.

I would like to thank my family for their support, and additionally thank my mother for proof-reading the text of this diploma. Also, I thank my partner for supporting me during late stages of this research.

Bc. Valeriia Iegorova

Název práce:
Rekonstrukce trojdimenzionálních modelů zvířat

Autor: Bc. Valeriia Iegorova

Studijní program: Computer Vision and Digital Image
Obor: Open Informatics
Druh práce: Master thesis

Vedoucí práce: Doc. Ing. Tomáš Pajdla, Ph.D.

Abstrakt: Tato diplomová práce se zabývá problematikou 3D rekonstrukce zvířat, která by mohla mít řadu praktických využití v zemědělství nebo počítačové animaci. Poskytuje ucelený přehled literatury a pojednává o existujících metodách 3D rekonstrukce a jejich vhodnosti pro danou úlohu. Následně demonstruje výsledky nejvhodnějších současných end-to-end rekonstrukčních pipeline a navrhuje novou metodu nazvanou *SMAL4V*, která tyto pipeline rozšiřuje pro videosekvence. K tomuto účelu také představuje novou datovou sadu, *AnimalKey*, na které lze 3D rekonstrukční modely trénovat nebo porovnávat, a zpřístupňuje ji výzkumné komunitě.

Klíčová slova: 3D rekonstrukce, neuronové sítě, strojové učení

Title:
Reconstructing 3D Models of Animals

Author: Bc. Valeriia Iegorova

Abstract: This diploma thesis explores the problem of 3D animal reconstruction, which could have multiple practical applications in agriculture or computer animation. It provides a comprehensive literature overview, discussing existing 3D reconstruction methods and their suitability for the given task. It then demonstrates the results of the most suitable state-of-the-art end-to-end reconstruction pipelines and proposes a new method, called *SMAL4V*, which extends those pipeline to video sequences. For that, it also introduces a new dataset, *AnimalKey*, on which 3D reconstruction models can be trained or compared, and makes it available to the research community.

Key words: 3D reconstruction, artificial neural networks, machine learning

Contents

List of Acronyms xi

List of Figures xii

1 Introduction 1
1.1 Motivation 1
1.2 Thesis structure 3
1.3 Contributions 3

2 Background Knowledge 5
2.1 Epipolar geometry 5
2.2 Non-rigidity and motion 8
2.3 Neural Networks 9
2.3.1 Convolutional Neural Networks 9
2.3.2 Recurrent Neural Networks 10
2.3.3 Neural Renderer 11
2.3.4 One-Shot Video Object Segmentation 13

3 Literature Overview 15
3.1 Overview of the Reconstruction Methods 15
3.1.1 Active reconstruction 16
3.1.2 Structure-from-Motion 19
3.1.3 Template-based methods 20
3.1.4 Model-based methods 21
3.2 Reconstruction Methods in Animal Modeling 23
3.3 Skinned Multi-Animal Linear Model 29
3.3.1 Global/Local Stitched Shape Model 29
3.3.2 Fitting to 3D scans 29
3.3.3 Fitting to images 30
3.4 End-to-End Learning Pipelines 30

4 Problem Statement 33

5 Methodology 35
5.1 Datasets 35
5.1.1 StanfordExtra 36
5.1.2 TigDog 36
5.1.3 Benchmark Animal Dataset of Joint Annotations 37
5.1.4 AnimalKey 37
5.2 Proposed Pipeline 38
5.2.1 Architecture of the test network 39
5.3 Training Procedure 41

6 Experimental Results 43
6.1 Preliminary Experiments 43
6.1.1 Reconstruction with rigid techniques 43
6.1.2 Reconstruction with RGB-D cameras 46
6.2 Experiments with Parametric Models 48
6.2.1 Experiments with the SMALST pipeline 48
6.2.2 Experiments with the WLDO pipeline 49
6.2.3 Comparison on AnimalKey and TigDog 49

7 Conclusions 53
7.1 Results 53
7.2 Future Work 54

Bibliography 55

List of Acronyms

ANNs	Artificial Neural Networks
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
DIFs	Deep Implicit Functions
RANSAC	RANdom SAmple Consensus
ARAP	As-Rigid-As-Possible
PCK	Percentage of Correct Keypoints
IoU	Intersection-over-Union
OSVOS	One-Shot Video Object Segmentation
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
ToF	Time-of-Flight
SfM	Structure-from-Motion
NRSfM	Non-Rigid Structure-from-Motion
SCAPE	Shape Completion and Animation for PEople
SMPL	Skinned Multi-Person Linear SMAL with Refinement
SMAL	Skinned Multi-Animal Linear
SMALR	SMAL with Refinement
SMALST	SMAL with learned Shape and Texture
SMBLD	Skinned Multi-Breed Linear Model for Dogs
SMAL4V	SMAL for Videos
CGAS	Creatures Great and SMAL
WLDO	Who Left the Dogs Out
BADJA	Benchmark Animal Dataset of Joint Annotations
GLoSS	Global/Local Stitched Shape
HMR	Human Mesh Recovery
VIBE	Video Inference for Body Pose and Shape Estimation

List of Figures

1.1 A typical example of the point clouds obtained from the reconstruction of some static objects. Image taken from [1]. 1

1.2 An example of the 3D reconstruction of a cow. A photogrammetry pipeline from RealityCapture [18] was used to create this example. . . 2

2.1 Geometry of the point correspondences. (a) A point in space X and its images x_1 and x_2, creating the epipolar plane . (b) A point in 2D image x_1, back-projected into 3D space. Its corresponding point as seen with the right camera would lie on the epipolar line l_2. Image taken from [23]. 5

2.2 A pair of images with a subset of detected inlier image correspondences and their epipolar lines. 6

2.3 Example of the automatically detected inlier point correspondences. The origin of each arrow corresponds an interest point in the given image, while the head of the error corresponds to the position of its corresponding point in the other image. 8

2.4 Example frames showing a jelly fish moving and deforming at the same time. Image taken from [27]. 8

2.5 Left: a fully-connected neural network. Right: a convolutional neural network. Image taken from https://cs231n.github.io/convolutional-networks/. 9

2.6 2D convolutions with different padding. Image taken from [32]. 9

2.7 Neural networks with different input and output types (with the left-most corresponding to a vanilla ANN and the rest corresponding to different versions of RNNs). Image taken from http://karpathy.github.io/2015/05/21/rnn-effectiveness/ 10

2.8 Left: LSTM hidden cell. Right: GRU hidden cell. Image taken from https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 11

2.9 Calculation of approximate derivatives when: a pixel is outside of the rendered face (left), pixel is inside of the rendered face (right). Image taken from [37]. 12

2.10 NeuralRenderer used to enable backpropagation through a simple 3D reconstruction pipeline. Image taken from [37]. 12

2.11 Example segmentation results by OSVOS. Image taken from [38]. . . 13

2.12 Overview of the OSVOS pipeline. Image taken from [38]. 13

3.1 Example from the Microsoft Kinect camera. Left: image from the infrared camera, with structured light pattern visible in the IR spectrum. Right: the corresponding depth map with color map going from white (nearest) to blue (furthest). Images taken from Wikipedia. 16
3.2 Examples of RGB-D cameras, based on the structured light technology. Left: ASUS Xtion Pro. Middle: Intel RealSense SR305. Right: Microsoft Kinect for Xbox 360. 17
3.3 Examples of RGB-D cameras. Left: active stereo cameras (Intel RealSense D415 and Intel RealSense D435). Right: Time-of-Flight (ToF) cameras (Microsoft Kinect v2, MESA SwissRanger 4000, Creative Senz3D, Soft Kinectic DS325). 17
3.4 Left: the input ToF scan. Color: the normal maps after applying KinectFusion. Grayscale: Phong-shaded renderings after KinectFusion. Image taken from [63]. 18
3.5 Real-time reconstruction of a dynamic scene with DynamicFusion. Image taken from [64]. 18
3.6 From left to right: input frame, reconstruction with DynamicFusion [64], reconstruction with DeepDeform [68], reconstruction with Neural Non-Rigid Tracking [67]. Image taken from [67]. 18
3.7 Left: examples of images of an object taken by a single moving RGB camera. Right: rigid object of interest and the camera positions (in red) from which the images were taken. Image taken from [70]. 19
3.8 Example of the template-based reconstruction. From left to right: (a) reference image with 2D-to-3D correspondences marked, (b) input template, (c) input RGB image, (d) resulting 3D reconstruction. Image taken from [55]. 20
3.9 Example of a human SMPL model. From left to right: (a) human template, (b) template with the custom shape deformations, (c) template with the additional pose-dependent deformations, (d) posed and deformed template. Image taken from [86]. 21
3.10 Top: example of the learned category-specific templates. Image taken from [94]. Bottom: example of a Skinned Multi-Person Linear (SMPL) model, clothed with the use of deep implicit functions Image taken from [94]. 22
3.11 Left: input RGB frames. Right: reconstruction of the giraffe's body surface. Image taken from [49]. 23
3.12 Top: input RGB frames and corresponding animal contours. Bottom: resulting 3D reconstructions. Image taken from [96]. 23
3.13 From left to right: (a) reference image with 2D-to-3D correspondences marked, (b) input template, (c) input RGB image, (d) resulting 3D reconstruction. Image taken from [55]. 24
3.14 Each left: input RGB image with user-clicked keypoints. Each right: the resulting 3D reconstruction. Image taken from [54]. 24
3.15 Left: input RGB images with corresponding contours of individual pre-defined body parts. Center: the reconstruction of individual parts. Right: the resulting 3D model. Image taken from [76]. 25
3.16 Comparison between the results of SMAL (left) and SMALR (middle) image fitting. Image taken from [97]. 26

3.17 From left to right: input images, predicted silhouettes, predicted keypoints, post-processed keypoints, predicted mesh and the same mesh from another view, produced by CGAS pipeline. Image taken from [101]. 26
3.18 A 3D reconstruction pipeline for birds 3D reconstruction from a single-view image RGB input. Image taken from [21]. 27
3.19 An overview of the hierarchical fitting procedure for birds reconstruction. Image taken from [100]. 27
3.20 Example results for a multi-species reconstruction network trained only with silhouettes supervision. Image taken from [105]. 28
3.21 Reconstruction examples for the differential deformation model. Image taken from [106]. 28
3.22 From left to right: a 3D template with colored surface points, predicted CSM and predicted mesh articulation from two views. Image taken from [110]. 28
3.23 Example of 3D scans, obtained from the animal figurines, which are then used to train the SMAL model. Image taken from [88]. 29
3.24 Example of the reconstructed animal shapes, fitted to the corresponding images. Image taken from [88]. 30
3.25 The SMALST modeling pipeline. Image taken from [98]. 31
3.26 The WLDO modeling pipeline. Image taken from [99]. 31
3.27 Top: synthetic zebra images, generated for the SMALST model training. Image taken from [98]. Bottom: real-world dog images from StanfordExtra, annotated for the SMBLD model training. Image taken from [99]. 32

5.1 Examples of images and their corresponding annotations from the StanfordExtra dataset. Image taken from [99]. 36
5.2 Examples of images and their corresponding annotations from the TigDog dataset. Image taken from [115]. 36
5.3 Examples of images and their corresponding annotations from the BADJA dataset. Image taken from [101]. 37
5.4 Examples of images from the AnimalKey dataset. 37
5.5 A screenshot of the labeling interface, prepared for the keypoints collection. 38
5.6 Overview of the proposed SMAL4V pipeline. Created with draw.io tool. Similarity with diagram design from [38] is intended to highlight the analogies between two approaches. 39
5.7 Proposed architecture of the SMAL4V's test network. Created with draw.io tool. Best viewed in color. White: available input, green: pre-trained from the parent network (dotted border indicates frozen weights), red: layers specific to the test network, blue: intermediate and final network outputs, yellow: individual loss terms, violet: models that are used as black-boxes in the proposed pipeline. 40

6.1 Example frames from a horse video, collected with an Apple iPhone 12 Pro. 44
6.2 Example frames from a horse video, collected with a Samsung S10. . 44
6.3 Output of a keypoint detector on a frame from a horse video. 44

6.4 Example frames from a cow video. 45
6.5 Top: 3D model obtained from Reality Capture on a nearly-static animal. Bottom: significantly decreased reconstruction quality once the animal slightly changed its pose between the frames. 45
6.6 3D model obtained from Reality Capture on: nearly-static animal (top), freely-moving animal (bottom). 46
6.7 3D data obtained with a HoloLens mixed-reality headset. 47
6.8 3D model obtained from Scaniverse on: nearly-static animal (top), freely-moving animal (bottom). 47
6.9 Example results of the SMALST pipeline on the Grevy Zebra's test set. From left to right: cropped input image, overlayed predicted mesh (predicted keypoints in red, ground-truth keypoints in black), predicted mask, ground-truth mask. 49
6.10 Example results of the SMALST pipeline on the AnimalKey's single-frame test set. From left to right: cropped input image, overlaid predicted mesh (predicted keypoints in red, ground-truth keypoints in black), predicted mask, ground-truth mask. 50
6.11 Example results of the WLDO pipeline on the AnimalKey's single-frame test set. From left to right: cropped input image, overlaid predicted mesh (predicted keypoints in red, ground-truth keypoints in black), predicted mask, ground-truth mask. 50
6.12 Example results of the SMALST pipeline on the TigDog's single-frame test set. From left to right: cropped input image, overlayed predicted mesh (predicted keypoints in red, ground-truth keypoints in black), predicted mask, ground-truth mask. 51
6.13 Example results of the WLDO pipeline on the TigDog's single-frame test set. From left to right: cropped input image, overlaid predicted mesh (predicted keypoints in red, ground-truth keypoints in black), predicted mask, ground-truth mask. 51

Chapter 1

Introduction

This thesis deals with the problem of 3D animal reconstruction and describes the issues that arise in modern reconstruction pipelines when applied to the articulated moving objects. It provides a comprehensive overview of the modern and historical reconstruction methods and discusses their applicability to the given task. Following the successes of the neural networks in multiple areas of computer vision, it explores an end-to-end learning-based family of the reconstruction pipelines in more details. Results of several such pipelines are demonstrated and a new method, called SMAL4V, is proposed, extending the existing end-to-end methods to the video sequences. The proposed architecture leverages temporal information to predict smoother meshes throughout the video.

Figure 1.1: A typical example of the point clouds obtained from the reconstruction of some static objects. Image taken from [1].

1.1 Motivation

The research of 3D reconstruction has been a popular topic in computer vision for several decades. It has a wide range of possible applications, including medicine [2], agriculture [3][4], archaeology [5], ecology [6], robotics [7], autonomous vehicles [8], motion capture [9] and many others. A typical 3D reconstruction pipeline takes a set of images of an object or scene of interest and returns its 3D representation, such as a point cloud, mesh or voxel model. Optionally, the resulting representation can be colored or texturized if required by the application. Depending on