The HealthSign Project: Vision and Objectives

D. Kosmopoulos
University of Patras
dkosmo@upatras.gr

A. Argyros
Foundation for Research and Technology - Hellas
argyros@ics.forth.gr

C. Theoharatos
Irida Labs S.A.
htheohar@iridalabs.gr

V. Lambropoulou
University of Patras
V.Lampropoulou@upatras.gr

C. Panagopoulos
Bioassist S.A.
cpan@bioassist.gr

I. Maglogiannis
University of Piraeus
imaglo@unipi.gr

ABSTRACT

This paper presents the HealthSign project, which deals with the problem of sign language recognition with focus on medical interaction scenarios. The deaf user will be able to communicate in his native sign language with a physician. The continuous signs will be translated to text and presented to the physician. Similarly, the speech will be recognized and presented as text to the deaf users. Two alternative versions of the system will be developed, one doing the recognition on a server, and another one doing the recognition on a mobile device.

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; • **Hardware** → *Integrated Circuits*;

KEYWORDS

sign language recognition, embedded processing

ACM Reference Format:
D. Kosmopoulos, A. Argyros, C. Theoharatos, V. Lambropoulou, C. Panagopoulos, and I. Maglogiannis. 2018. The HealthSign Project: Vision and Objectives. In *PETRA '18: The 11th PErvasive Technologies Related to Assistive Environments Conference, June 26–29, 2018, Corfu, Greece.* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.1145/3197768.3201547

1 INTRODUCTION

Sign Languages (SLs) are the main means of communication for deaf people. The access to SL is essential for the fulfillment of basic Human Rights, however there is a shortage of interpreters, which undermines these rights and often endangers the lives of the deaf, especially in cases of emergency or serious health incidents.

HealthSign proposes to develop an application for the automated interpretation of the Greek Sign Language (GSL) over internet with focus on the health services, which are the most common reason to seek for an interpreter. The high demand for interpreters is often not met or requires long waiting. On the other hand, the availability of interpreters facilitates the integration of the deaf community into the society and .

Vision is probably the only sensor modality that could be of practical use because (a) only vision can capture manual and non-manual cues, which provide essential information for SLR, (b) camera-equipped hand-held devices with powerful processors are a commodity nowadays and (c) recent advances in computer vision and machine learning render mainstream visual SLR a realistic option.

The HealthSign project aims to fulfill the following innovative goals:

- Develop a database of GSL from native speakers with emphasis on health services.
- Implement an internet-based platform for synchronous communication and interpretation with health professionals.
- Develop in parallel a lightweight version which will be able to run on an embedded platform.
- Develop algorithms for recognition of SLs, using computer vision and deep learning using the hand and body/facial cues.
- Implement the algorithms on embedded platforms using FPGAs.

In the long term we aim at the viability of the proposed application, which will be achieved by (a) simple off-the-shelf equipment for the users, (b) efficient implementation of the proposed algorithms, (c) simple installation, and (d) development of a business plan to facilitate the longevity of the proposed product.

The consortium is composed of (a) The Signal Processing and Telecommunications Lab of the University of Patras as expert in machine learning, which is necessary for the recognition of GSL, (b) the Computer Vision and Robotics Lab of ICS-FORTH, which will adapt their 3D hand model for tracking, (c) the Bioassist company (consulting from University of Piraeus), which specializes in assistive technologies and develops an internet based platform, (d) the IRIDA S.A. to develop the embedded application, and (e) the Deaf Studies Unit at the University of Patras, which will bring the users, the interpreters and the GSL experts.

In the next section we describe the detection and tracking methods we are going to use. Section 3 describes the methods for sign language modeling and recognition, while section 4 presents the architecture for SLR on the server. The section 5 describes how we plan to do SLR on a mobile device and section 6 concludes this paper.

PETRA '18, June 26–29, 2018, Corfu, Greece

ACM ISBN 978-1-4503-6390-7/18/06…$15.00
https://doi.org/10.1145/3197768.3201547

2 DETECTION AND TRACKING

The problem of detection and tracking of human body and hands using visual sensors have been studied a lot in the past, e.g., [10], [11], [17]. Apart from the theoretical interest, several such solutions can be employed for human-machine interfaces, such as video games, virtual and enhanced reality. One of the most interesting such applications is the sign language understanding, since the 3D motion of human body and hands can feed higher level processes.

The methods to estimate hand poses can be classified as discriminative and generative as described in [13]. The former ones include rather computationally intensive methods that learn a mapping from visual data to hand poses and then apply this mapping to future input. Their fundamental problem is that their accuracy is limited, since they assume learning a predefined, and thus limited limited, set of poses. In the generative methods typically a geometric model is used to infer the optimal pose parameters. Despite the increased computational cost these approaches offer higher accuracy if enough computational resources are available.

Despite the progress in the last decade there is still room for significant improvements. The main issue for tracking is the speed and dexterity of the hand motion [5]. The practical problems include the initialization and the re-initialization of tracking in case of failure.

Here we propose the adaptation and improvement of the state of the art ([7], [10]) for detection and tracking of the upper body and the hands to use in sign language recognition. This is not trivial, due to the high requirements for accuracy and usability.

We will use sequences of hand and body poses using color cameras and depth sensors (RGBD) or alternatively stereo cameras. We will record off-line predefined gestures and we will estimate the poses using color and depth. The poses along with the images will be used in the next learning steps.

In our attempt to improve our method [10] we will use higher resolution data, stereo images, as well as methods for the detection of hand-tips. To this end we will use the Kinect-2 sensors which offer higher image resolution and frame rate, high rate cameras or narrow baseline stereo cameras with high frame rate. We will also develop an initialization and reinitialization procedure using discrete methods. Such a method can be based on neural networks for the direct estimation of the pose. The availability of additional information such as the position of the hand-tips can contribute to initialization accuracy and better tracking. The goal is to learn the function that associates the body/hand pose through depth only for offline pose estimation. The online association of image to pose will have the form of a probability density function.

The goal is to be able to follow about 200 gestures and their variations. For the offline pose estimation we aim at a rate of about 10 frames per second.

3 SIGN LANGUAGE MODELING AND RECOGNITION

Typically, generative or discriminative methods under a time-series classification framework have been used (e.g., [1],[16]). For large-scale SLR, researchers have focused on the essential linguistic parameters involved in sign production, including hand configuration, orientation and trajectory [4]. Hand shape recognition performed in 2D can be enhanced by exploitation of linguistic constraints [3]. However, accuracy is severely limited by the inability of a 2D model to capture the entire manifold of the 3D pose. On the other hand, gesture recognition systems in 3D rarely consider the linguistic constraints posed by the language to recognize. Furthermore, the importance of non-manual cues (facial expressions, full body motion) should not be ignored, since they can improve accuracy. Taking into consideration this kind of information gives rise to additional feature streams, the fusion of which with manual ones can yield much more effective SLR systems compared to the current state-of-the-art. Indeed, an end-to-end trained system that can effectively exploit the available visual information sources, as well as the linguistic and physical constraints, is expected to alleviate the need of large datasets for its effective training, thus giving rise to realistic applications. Finally, an additional challenge is to enable real-time operation of the system, under the constraint of limited (inexpensive) availability of computational resources.

Convolutional neural networks (CNNs) could be the basis for such a system since they yield state-of-the-art performance in benchmark datasets, as opposed to hand-crafted features[12]. Deep feature extractors were applied on sparse coding (SC), deep belief networks, and stacked autoencoders. They result in highly-informative coarse-to-fine representations, which can be utilized in classification pipelines. However, such pipelines are essentially treated as black-box solutions with ad-hoc architectures, that lack a solid rationale based on a well-defined mathematical background.

We will do the modeling by combining (a) deep networks (b) the tracking results and (c) linguistic constraints. The deep networks provide state of the art performance and the linguistic constraints along with the tracking results are expected to drive the optimization close to optimal solutions.

An initial approach will employ available networks such as the AlexNet [6] or the GoogLeNet [14], which have been trained with ImageNet. The output sequences will be classified with a conditional random field using the L-BFGS algorithm [9]. We will experiment on how to approximate of the Hessian matrix.

Our second approach will use data from body/hand tracking without using depth. This is because (a) we cannot expect the mainstream devices to have depth sensors and (b) depth is not really necessary, but only for training. From tracking we will get the conditional probability of the pose given the image. The pose can be used for sign language recognition with distance matrix regression [8], which offers some advantages like: (a) detailed generative model which renders the system more tractable (b) insertion of linguistic constraints like the initial/final hand/body/face configurations, which may be easily integrated as Bayesian priors.

4 SIGN LANGUAGE RECOGNITION ON SERVER

Recently, several ideas for products have been presented, e.g., based on multiple vision sensors, gloves and wristbands that measure electrical activity by muscles. Unfortunately, they all remain prototypes, which cannot be widely employed for Sign Language Recognition (SLR), due to the size and the obtrusiveness of the required equipment. Vision is probably the only sensor modality that could be of practical use because due to capturing manual and non-manual

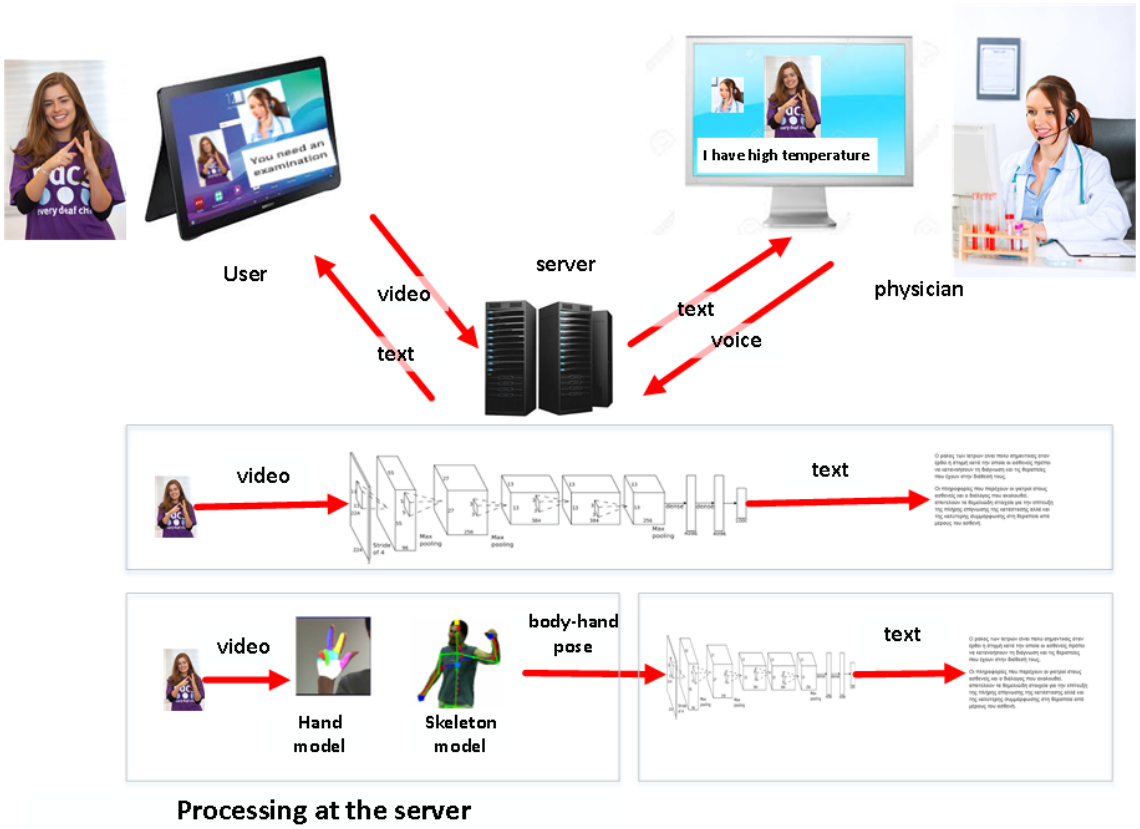

Figure 1: The HealthSign architecture for mobile devices with the processing on the server. The processing consists in the direct interpretation of the video which is transmitted from the user device. It is then transformed to text directly or after the extraction of the hand/body pose. The physician's speech is translated to text on the server side using a speech recognition software and then transmitted to the deaf user.

cues (e.g., facial expressions), due to the availability of cameras and powerful processors on most mobile devices and due to recent advances in computer vision and machine learning render mainstream visual SLR a realistic option.

The proposed platform will integrate the software tools to be developed as described in the previous sections (Fig. 1). It will also include a user interface (a) for the deaf user, to display the physician and their responses as text (using a commercial speech recognition software) and (b) for the physician to display the deaf user and the related signs translated to text, or as voice using a text-to-speech software.

According to the use scenario a deaf user feels sick and decides to consult a physician through the platform. He has installed the application on his tablet/smart phone. By pressing a button he connects to the application and is able to start the discussion, while the camera faces him. The video frames are transmitted to the server where the interpretation is made. The sign language is interpreted to text and the text is transformed to speech by a text-to-speech tool. The audio is sent to the physician. The physician talks to the patient and the speech data are transmitted to the server, where they are transformed to text. The text is sent to the deaf's client device and appear on its screen.

The internet communication platform will support the transmission of video via WebRTC on android devices. It will support the interface parametrization, as well as a lot of web services. It has been tested as a Bioassist product (Heart Around) [2]. It will be modified to support the aforementioned functionalities. The goal is to perform the whole cycle (recording-interpretation-presentation) in less than 5 seconds.

The experiments will require data collection, which will be done by using native deaf users and GSL interpreters. Several hundreds hours of videos will be required to cover the most interesting scenarios. The data collection will be done using RGB and depth cameras (or stereo cameras) under realistic conditions, i.e., continuous sign language under visual noise. About two hundred common phrases will be modeled to allow for a basic communication. Related public databases will also be exploited at the beginning of the project.

There are several concerns that have not been examined, such as confidentiality and privacy issues: the transmitted information is very sensitive but how do we ensure that it remains confidential? What are the limits between usability and privacy in such a system?

What information should be stored and what should be discarded for better safety? These will be analyzed during the user requirements capturing phase, having in mind the international best practices.

5 EMBEDDED SIGN LANGUAGE RECOGNITION ON MOBILE DEVICES

We propose the development of SLR on platforms like Qualcomm Snapdragon or NVidia GPU for local real time processing on mobile devices. The system is described in Fig. 2. The implementation of deep learning networks on embedded systems faces a lot of challenges. These networks are quiet complex and require a lot of processing power for real time tasks. There are two main approaches (a) efficient programming model and (b) simplified networks.

In that scenario the video frames are not transmitted to the server, but are processed locally. The sign language is interpreted to text and the text is sent to the server where it becomes transformed to speech by a text-to-speech tool. The audio is sent to the physician. The physician talks to the patient and the speech data are transmitted to the server, where they are transformed to text. The text is sent to the deaf's clinet device and appear on its screen.

The state of the art embedded systems offer System-On-Chip with high processing power, which is distributed on several units i.e., GPU, CPU, DSP. They can include several processing units in different levels, e.g., in Qualcomm SnapDragon or in Samsung Exynos the CPU is composed of several cores (eight). Every core is composed of subunits that require separate programming (e.g., ARM/NEON). To employ all available computational resources heterogeneous programming must be employed for the different units (C/C++, ARM intrinsics, Assembly for the CPU, OpenCL for the GPU etc). This type of programming requires knowledge of each unit, optimized code and unit synchronization.

The aforementioned method often is not enough and the computation reaches its limits. Therefore the simplification of the networks is necessary. In deep learning there is the problem of finding the appropriate network size. Often the networks are larger than required. To this end IRIDA has developed the innovative technique of parsimonious inference [15]. According to it a network is modified with the addition of specific units which regulate the function of computational units (convolutional kernels) in real time. For a specific image only some of them are active, for another image some others etc. Therefore a significant amount of resources is saved.

6 CONCLUSIONS

We have introduced the basic concepts behind the HealthSign project. We have presented the basic elements of the proposed architecture and the principles of their implementation. We presented the two alternative architectures that we are going to develop, the first doing the processing on the server and the other one doing the processing on the mobile device. In the near future we are going to begin the implementation and the experimentation with real users. There are a lot of challenges to deal with, the most obvious being the modeling of the signs in a continuous form and the real time operation.

ACKNOWLEDGMENTS

This work is partially supported by the Greek Secretariat for Research and Technology, and the EU, Project HealthSign: Analysis of Sign Language on mobile devices with focus on health services T1EΔK-01299 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020.

REFERENCES

[1] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. 2009. A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 9 (Sept 2009), 1685–1699. https://doi.org/10.1109/TPAMI.2008.203

[2] Bioassist. 2018. Heart around. (2018). https://heartaround.com/?lang=el

[3] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. 2014. A New Framework for Sign Language Recognition Based on 3D Handshape Identification and Linguistic Modeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (26-31), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland.

[4] Liya Ding and Aleix M. Martinez. 2009. Modelling and Recognition of the Linguistic Components in American Sign Language. *Image Vision Comput.* 27, 12 (Nov. 2009), 1826–1844. https://doi.org/10.1016/j.imavis.2009.02.005

[5] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding* 108, 1 (2007), 52 – 73. https://doi.org/10.1016/j.cviu.2006.10.012 Special Issue on Vision for Human-Computer Interaction.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257

[7] Damien Michel, Ammar Qammaz, and Antonis A. Argyros. 2017. Markerless 3D Human Pose Estimation and Tracking Based on RGBD Cameras: An Experimental Evaluation. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '17)*. ACM, New York, NY, USA, 115–122. https://doi.org/10.1145/3056540.3056543

[8] Francesc Moreno-Noguer. 2017. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 1561–1570. https://doi.org/10.1109/CVPR.2017.170

[9] Jorge Nocedal. 1980. Updating Quasi-Newton Matrices with Limited Storage. *Math. Comp.* 35, 151 (1980), 773–782. http://www.jstor.org/stable/2006193

[10] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. 2011. Efficient model-based 3D tracking of hand articulations using Kinect. In *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*. 1–11. https://doi.org/10.5244/C.25.101

[11] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. 2014. Realtime and Robust Hand Tracking from Depth. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1106–1113. https://doi.org/10.1109/CVPR.2014.145

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (01 Dec 2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[13] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. 2015. Accurate, Robust, and Flexible Real-time Hand Tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3633–3642. https://doi.org/10.1145/2702123.2702179

[14] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[15] Ilias Theodorakopoulos, V. Pothos, Dimitris Kastaniotis, and Nikos Fragoulis. 2017. Parsimonious Inference on Convolutional Neural Networks: Learning and applying on-line kernel activation rules. *CoRR* abs/1701.05221 (2017). arXiv:1701.05221 http://arxiv.org/abs/1701.05221

[16] Christian Vogler and Dimitris Metaxas. 2004. Handshapes and Movements: Multiple-Channel American Sign Language Recognition. In *Gesture-Based Communication in Human-Computer Interaction*, Antonio Camurri and Gualtiero

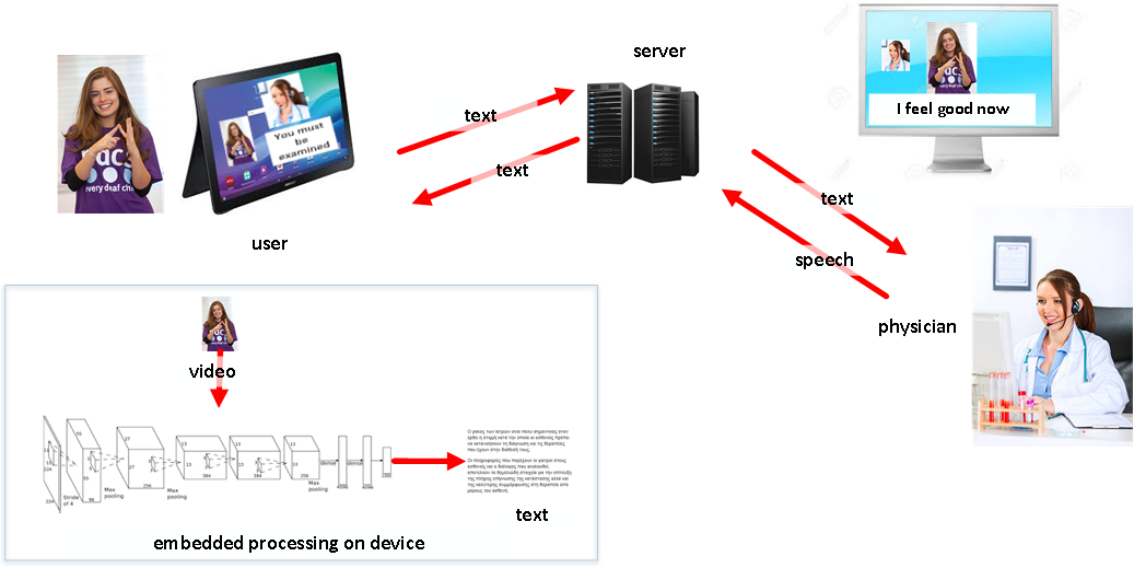

Figure 2: The HealthSign architecture for mobile devices with the processing on the device. The interpretation is effected on the device and the results are transmitted as text. The rest is similar to Fig. 1.

Volpe (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 247–258.

[17] Qi Ye, Shanxin Yuan, and Tae-Kyun Kim. 2016. Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 346–361.