FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 33, N° 3, September 2020, pp. 379-394
https://doi.org/10.2298/FUEE2003379G

COMPARATIVE EVALUATION OF KEYPOINT DETECTORS FOR 3D DIGITAL AVATAR RECONSTRUCTION

Dušan Gajić, Gorana Gojić, Dinu Dragan, Veljko Petrović

University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

Abstract. *Three-dimensional personalized human avatars have been successfully utilized in shopping, entertainment, education, and health applications. However, it is still a challenging task to obtain both a complete and highly detailed avatar automatically. One approach is to use general-purpose, photogrammetry-based algorithms on a series of overlapping images of the person. We argue that the quality of avatar reconstruction can be increased by modifying parts of the photogrammetry-based algorithm pipeline to be more specifically tailored to the human body shape. In this context, we perform an extensive, standalone evaluation of eleven algorithms for keypoint detection, which is the first phase of the photogrammetry-based reconstruction pipeline. We include well established, patented Distinctive image features from scale-invariant keypoints (SIFT) and Speeded up robust features (SURF) detection algorithms as a baseline since they are widely incorporated into photogrammetry-based software. All experiments are conducted on a dataset of 378 images of human body captured in a controlled, multi-view stereo setup. Our findings are that binary detectors highly outperform commonly used SIFT-like detectors in the avatar reconstruction task, both in terms of detection speed and in number of detected keypoints.*

Key words: *Detector, Photogrammetry-based reconstruction, 3D human avatar, Structure from Motion, Multi-view Stereo*

1. INTRODUCTION

An avatar is a digital self-representation of a participant in a computer generated virtual world [1] and can be represented both in two (2D) or three dimensions (3D). The significance of 3D avatars is constantly growing due to the expansion of virtual worlds in which participants identify themselves through their avatars. Recently, avatars have been successfully involved in many applications, including entertainment [2], shopping [3], education [4], health [5], and military [6].

For some applications, the avatar must be a 3D, highly personalized representation of a person, e.g., avatars used for meeting events or virtual try-on applications [3], [7]. Since it is a labor-intensive task to produce high-quality 3D avatars manually, many techniques for automatic generation have been proposed. One of them is digital photogrammetry

Received October 13, 2019; received in revised form January 12, 2020
Corresponding author: Dušan Gajić
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21102 Novi Sad, Serbia
E-mail: dulegajic@gmail.com

© 2020 by University of Niš, Serbia | Creative Commons License: CC BY-NC-ND

which is the subject of the research described in this paper. To obtain a 3D avatar through digital photogrammetry software, a series of overlapping images showing a person from different viewpoints are first acquired. A typical photogrammetry-based pipeline consists of three phases [8]: Structure from Motion (SFM), Multi-view Stereo (MVS), and mesh creation. As an input, the SFM phase receives a series of overlapping 2D images and outputs a 3D sparse point cloud. This phase relies on a triangulation process to recover 3D points from multiple 2D projections of the same 3D point present on two or more images. To identify 2D points on images that represent the same 3D point, an algorithm for point detection, in literature also known as a detector, is applied to all input images. This helps locate keypoints—patches of the image which represent the 3D points that will make up the sparse point cloud. Depending on type, detectors find keypoints corresponding to structures known as edges, blobs, or corners. Detected keypoints are matched with each other to find tracks of keypoints that represent the same 3D point using a description-generating algorithm. For more information about SFM, MVS, and mesh creation phases of the photogrammetry-based pipeline, we refer the reader to [8].

Recently, large number of detection algorithms have been proposed. Although minor discrepancies in the research on the evaluation of detection algorithms exist, scale-invariant feature transform (SIFT) based algorithms are still considered to be state-of-the-art algorithms for general-purpose use. However, it has been shown that even cutting-edge commercial software solutions that use SIFT or SIFT-like algorithms in phases of a keypoint detection, such as AgiSoft PhotoScan [9], have difficulties when reconstructing human avatars. Those difficulties are often caused by an insufficient number of detected keypoints on particular problem areas (e.g., backs), and ultimately result in an incomplete avatar model. According to [10], the optimal choice of detector might depend on properties of the input data. This means that SIFT and SURF might not perform best in the specific case of 3D avatar reconstruction. Additionally, the price of software used for avatar reconstruction could be reduced if a patent-free detector algorithm were to be used. Recently, patent-free detectors have been implemented in some of the leading open-source photogrammetry-based solutions, such as Meshroom [11] and OpenMVG [12]. Many of the detectors tested in our study have been proposed considerably after SIFT and SURF algorithms, thus it is expected that more of them will be implemented in photogrammetry software in the future to compensate for shortcomings of SIFT and SURF. All detectors included in this study, including those already incorporated into photogrammetry software, can be used for human avatar reconstruction. Still, it remains a question if detectors not yet implemented in available photogrammetry software could yield a comparable or better result to those already implemented. From this viewpoint, our study can be seen as a first step to guide the implementation of human-based photogrammetry software.

To this end, we have conducted an extensive, standalone detector evaluation study on a human-based image dataset captured in controlled conditions. The results of such a study can lead to less expensive, more widely-available photogrammetry software, if it shows that free-to-use detection algorithms can replace SIFT without sacrificing quality. We evaluate eleven detectors, both binary and floating-point in terms of the number of keypoints detected, detection speed and detector efficiency in finding keypoints in the region of an image representing a person. Our overall findings are that binary detectors highly outperform floating-point detectors tested in this study, including SIFT and SURF detectors, for the task of 3D human body reconstruction.

The rest of the paper is organized as follows. In Section 2, we present a brief overview of work in the field of detector evaluation. Section 3 discusses in detail the experimental framework for evaluation, as well as the dataset used in the experiments. We present and discuss the obtained results in Section 4. The final section offers the main conclusions, as well as possible directions for future work.

2. Related Work

In this section, we give a brief overview of the work related to human body reconstruction. We start by presenting associated studies in the detector evaluation field and report established state-of-the-art results. Next, we discuss different techniques for 3D reconstruction of a clothed human body representing an avatar, with a particular interest in model's level of detail.

2.1. Detector Evaluation

Detector evaluation has been a widely addressed topic in a computer vision. Extensive standalone detector evaluation for the use case similar to ours have been proposed in [13]. Ten well-established detectors at the time the paper was written were evaluated on a dataset captured in a multi-view stereo setup showing complex, non-planar scenes such as buildings, fruits, etc. Authors evaluate detectors through three metrics: recall rate introduced in [14], keypoint location, and the average number of detected keypoints. To calculate the first two metrics, they use a ground-truth data in the form of known camera positions and a 3D dense point cloud of a scene captured by a laser scanner. Since we do not have a precomputed 3D model of a person that can be used for recall and location calculation, we adopt the average number of keypoints as a metric in our work. Results of the experiments conducted in [13] show that FAST (Features from accelerated segment test) detector showed unreliable performance despite the large average number of detected keypoints. Although not so extensive, one of the most influencing works in the field of detector evaluation is the early work of Mikolajczyk and Schmid [15]. To evaluate the performance of the detectors under an extensive set of image transformations, the authors used ground-truth homographies between image pairs to match detected keypoints. This solution for keypoint verification is commonly used in experiments performed on images showing planar scenes, which is not valid in our case, since the scenes used for detector evaluation are non-planar.

Along with standalone detector evaluation, recent studies provide detector evaluation jointly with description algorithms through feature matching task [10], [16]–[18]. Joint detector-descriptor evaluation has been appealing due to the nature of the keypoint matching problem. Keypoint matching between two images is a two-step problem: (1) all keypoints are detected on both images and (2) described by a descriptor algorithm of the choice. Then, keypoint pairs are tracked by similarity in terms of descriptor's output. However, introducing descriptors into detector evaluation adds more complexity to the evaluation task, since the final performance cannot be assigned solely to detection or description algorithm, but rather to the combination of these two. In [16] SIFT, SURF, MSER (Maximally stable color regions), FAST and ORB (Oriented FAST and rotated BRIEF) detectors are evaluated in terms of fast matching on a dataset with different geometric and photogrammetric transformations including rotation, scale change, viewpoint

change, image blur, JPEG compression and change in illumination. In [17] more detectors are added in evaluation, including CENSURE, AGAST (Adaptive and generic accelerated segment test), and BRISK (Binary robust invariant scalable keypoints) over the extensive image transformations dataset comprised of multiple well-known feature evaluation datasets. Although commonly employed metrics for joint detector evaluations include repeatability score, precision and recall value, number of keypoint correspondences and keypoint detection time, in our experiment we adopt just the keypoint detection time since the other metrics are descriptor dependent.

There is a majority consent between proposed evaluation methods that FAST is one of the top-performing detectors in terms of the number of detected keypoints and detection speed. Considering the detection speed, FAST is followed by other binary detectors such as ORB and AGAST [17]. Although FAST expresses superior performance when it comes to the number of keypoints detected, it is stated in [13] that it was unreliable compared to the other scale-space keypoint detectors, such as Difference of Gaussian (DoG), today incorporated as a part of SIFT detector. According to a ranking proposed in [19], the best performing detector-descriptor combinations were FAST+SIFT and FAST+BRISK. In [20], a novel method for detector evaluation is introduced through the reconstruction of a 3D dense point cloud. Although authors compare just SIFT and AKAZE detectors, the method can be applied to other detection algorithms to verify already produced numerical results additionally. As future work, we intend to incorporate a similar approach in our evaluation framework.

2.2. Human Body Reconstruction

To reconstruct the 3D body model of a clothed human, affordable image-based techniques are used as an alternative to more expensive laser scan and structured light techniques [21]. Image-based reconstruction requires one [22]–[25] or more [26]–[29] temporally [26][30] or spatially [27] connected images captured by RGB [22]–[25] or RGB-D [26], [27], [30] sensors. Early work in this field was directed toward general-purpose multi-view stereo algorithms. In a multi-view stereo, multiple sensors are used in a setup to simultaneously capture images of the subject from different viewpoints with certain redundancy between the views. Although these techniques are not primarily designed for human body reconstruction, it has been demonstrated that highly detailed models can be obtained using this method [28], [29], [31], [32]. By design, multi-view stereo algorithms are sensitive to complex occlusions between the views, as well as sparse or repeated textures [28], [33], [34]. These appearances are ubiquitous in human body reconstruction: texture issues are often caused by clothes, and occlusions by nontrivial body shape and pose. As a result, the output body model may be missing some of the body parts [33], [34]. In [35], authors minimize model incompleteness by increasing redundancy between the views in a dense multi-view stereo setup. However, using tens or thousands of sensors in a setup significantly limits the proposed method's applicability due to high setup price and increased reconstruction time. Different approach to address the model incompleteness problem based on compressive sensing technique (CS) is presented in [36]. Compressive sensing has already been used to refine depth maps that are generated in later steps of the human avatar 3D reconstruction pipeline. This technique could be used to reduce the number of sensors in the setup, with a limitation that this approach can be applied just in cases where exact sensor positions are known during the image acquisition process, which is not the assumption in this paper. Still, it could be possible to apply CS to fill

missing parts of the final model reconstructed by the photogrammetry-based pipeline. It remains to be tested if CS could recover whole body parts or just minor patches on the model. Another effort to reduce the setup price was attempted in [27] where more affordable, low-resolution RGB-D sensors are used instead of RGB sensors. Although RGB-D sensors improve the reconstruction process results in terms of improved depth estimation, due to low sensor resolution, body models reconstructed in these setups lack details. Another approach to reduce the setup price is to use a sparser sensor setup. This is approach we utilize in our work. Since the redundancy between views in a sparse setup is low, models generated using these setups are more likely to be incomplete. To overcome this problem, algorithms for reconstruction from sparse setups usually do not rely solely on input images. In [24], [26] coarse human body template is used as a basis to overcome model incompleteness issues. The template is further modified according to input images to obtain a personalized model of a clothed person. The main disadvantage of using the template in the reconstruction process is unavailability to generate models with a high level of details.

Lately, human body reconstruction from a single image has been a topic of great interest in a computer vision. The most successful single-image approaches are those based on convolutional neural networks (CNNs) [22]–[25]. There are two common approaches to reconstruct a body model when it comes to CNNs: (1) estimate human body template parameters [37], [38] or (2) directly output voxel occupancy in the form of a voxel grid [22], [23]. The latter approach is of more interest to this work since it is more suitable for the reconstruction of a clothed body model. Recently, not just input color images, but also segmentation masks and body landmarks are used to output the clothed body model successfully. However, although voxel grid-based CNN reconstruction methods output promising results both in terms of completeness and level of details, this approach is currently limited by computational power to a voxel grid of approximate size of 128×128×128 voxels. This constraint is related to model detail level, which is limited by the maximal size of the grid.

It is of particular interest to our work that the 3D model of a clothed human body is highly detailed and complete. Thus, we choose multi-view setup with RGB cameras to capture images of a clothed subject since currently no other method can produce models with comparable high level of details. As a basis for our research, we use a general-purpose multi-view stereo reconstruction algorithm to obtain the clothed body model. Differently from the other work, we make an effort towards modifying general-purpose photogrammetry-based reconstruction algorithms for a human body reconstruction domain. To achieve that, we perform an extensive study of detector algorithms that are used as a first step in the pipeline to choose the best performing detectors on a human-based dataset. In this way, we tackle the problem of improving the reconstruction detail level and completeness through the improvement of the algorithm, instead of the more expensive sensor setup densification.

3. Experimental Framework

In this section, we explain the sensor setup used to capture the human-body based dataset used for the experiments, as well as a detailed description of conducted experiments. To refer to the person who has been photographed, we use a term *the subject*.

3.1. Camera Setup

To capture image data, we use multi-view stereo setup with 54 high-resolution RGB calibrated cameras, conceptually similar to the one described in [39] and shown in Fig. 2. During the image acquisition process, the subject is standing in a center of the setup with legs slightly apart and arms positioned at an approximately 30-degree angle away from the body (so-called A-pose) [39]. Fig. 1 gives an idea of the body areas visible on images captured by different cameras in the setup. Due to privacy concerns, we display subject silhouettes instead of color images. Almost all parts of the subject's body are visible on the captured images. The subject soles are the exception since they are not visible during the image acquisition process.

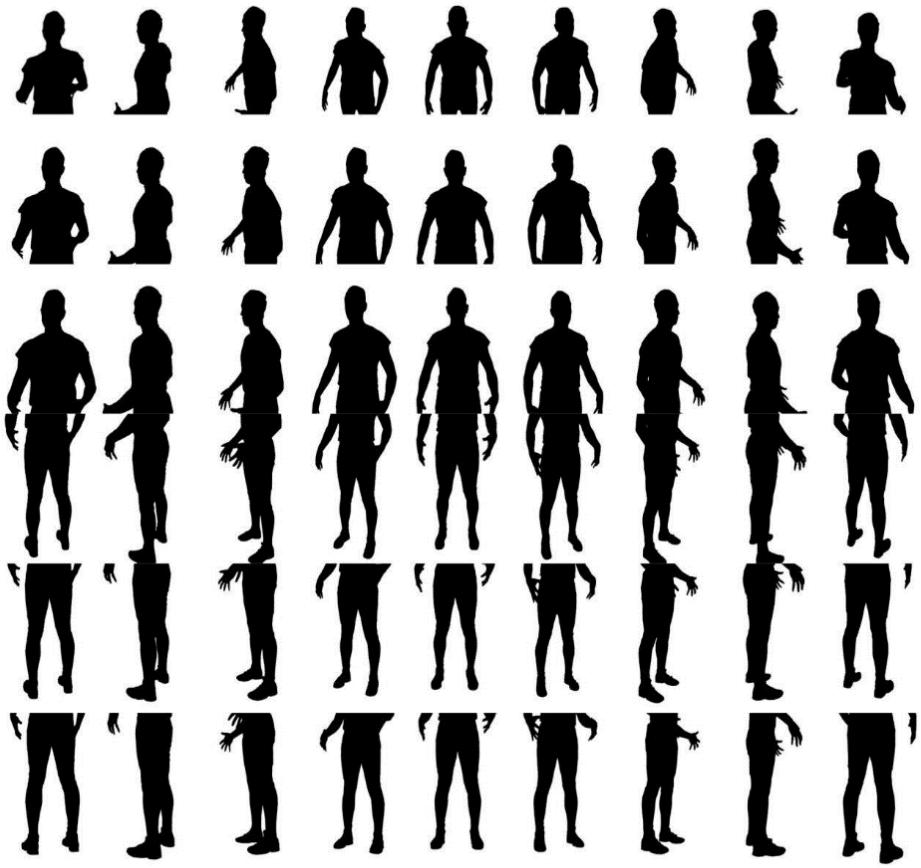


Fig. 1 Body coverage schema—anonymized real data

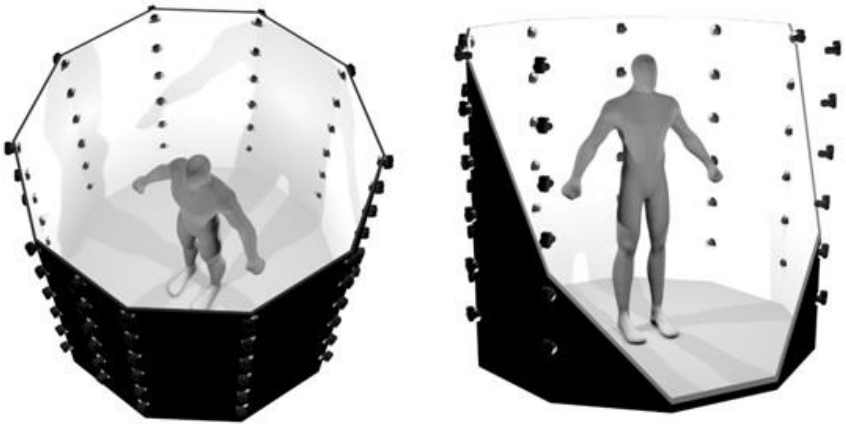
3.2. Dataset

The dataset we conduct experiments on consists of seven image sets that we will refer to as scans. Each scan is captured with the setup similar to [39] and consist of images displaying different body parts as illustrated in Fig. 1. To capture scans, we use two

different camera types (see Table 1). Due to relatively sparse camera setup, redundancy between images of a single scan is low. Images are captured from different viewpoints without precisely known camera positions. Some of the images may suffer from an illumination effect. Other frequently tested geometric or photogrammetric transformations in the general-purpose evaluations, such as rotation, blur or JPEG compression are omitted from the dataset since the presence of those transformations indicates an error in the scan acquisition process.

Table 1 Scans specification

Scan Identifier	Camera Manufacturer	Resolution
1, 2, 3, 4, 5, 6	Canon	3456x5184
7	Raspberry Pi	2464x3280



(a) (b)

Fig. 2 Conceptual camera setup shown from the top (a) and side view (b). Acquisition cameras are represented as columns of dark spots. This image is taken from [36]. In our work, we use acquisition setup similar to the one presented in the image.

3.3. Software and Hardware

All experiments are conducted on a personal computer running Windows 10 64-bit, powered by Intel i5-6600 CPU at 3.3 GHz, 32 GBs of RAM, and Nvidia GeForce 1050Ti graphics card. Evaluation pipeline is implemented as a single-threaded application using C++ programming language and compiled with Visual C++ 2015 compiler using speed optimizations (/O2 compiler flag). To make experiments easily reproducible, all detector implementations used are part of a publicly available OpenCV 3.2 [40] library. We compile the library from source with *opencv-contrib* package to include support for patented algorithms such as SIFT and SURF.

3.4. Detector Evaluation

We include both floating-point and binary detection algorithms into our study. As for floating-point algorithms we include currently the most popular SIFT [41] and SURF [42] detection algorithms, as well as STAR [43], Maximal self-dissimilarities (MSD) [44], Maximally stable color regions (MSER) [45], [46], and Good features to track (GFFT) [47]. We also evaluate number of binary detection algorithms such as Oriented FAST and rotated BRIEF (ORB) [48], Features from accelerated segment test (FAST) [49], Binary robust invariant scalable keypoints (BRISK) [50], Adaptive and generic accelerated segment test (AGAST) [51], and Accelerated KAZE (AKAZE) [52]. We choose to include in our study as many detectors as possible limiting ourselves to implementations available as a part of OpenCV library. When instantiating a detector object, we use default parameters for all detectors except for ORB and GFFT for which the maximum number of keypoints has been set to 300000 instead of much smaller, default values of 500 and 1000, respectively. We experimentally choose 300000 as an upper limit for the number of detected keypoints, since none of our test images exceed this limit under any detector.

We evaluate detectors on scans from Table 1. Experiments are conducted both on original scans and scans with a removed background (so-called masked scans). To remove the background, we apply a mask image to each image from the scan. A mask is new, binary image that corresponds to the original, color image with white pixels representing subject body and black pixels representing the background. Each mask image is manually labeled to precisely follow subject's outline. After the mask is applied, the image is left showing just the subject while the background is made entirely white. Since our study is also directed toward setup cost reduction, we are also interested in detectors performance in lower resolution images, since low-resolution cameras are cheaper. Motivated by this fact, we test all detector algorithms on images with applied scale factors of 1, 2, 4, and 8. To downscale original images, we use bilinear interpolation.

3.4.1. Performance metric

We use three metrics for the measurement of detector performance:

- *The average number of detected keypoints* has been calculated for both masked and original images. This metric is important for detector evaluation since the insufficient number of detected keypoints in the image segment showing the subject will almost certainly result in a sparse point cloud with too few points and, consequently, an incomplete avatar reconstruction. Certain areas of the human body, such as back or legs, can be particularly challenging to reconstruct due to a lack of edges or textures, which are detected as features by some detection algorithms.
- *The average number of keypoints per second = number of detected keypoints / time to detect keypoints*. Large number of keypoints is necessary to reconstruct complete and detailed avatar of a human, making keypoint detection time significant factor in a 3D reconstruction process. Choosing detector with large execution time might limit applicability of avatar reconstruction to non-realtime applications. Thus, similar to [19], we include detector execution time measurement into ours study. We improve the approach from [19], by not limiting the maximum number of keypoints detected by the algorithm. Since the number of keypoints detected by different detectors on the same image can vary, we do not measure absolute execution time as

introduced in [19]. Instead, we measure a detector's execution time indirectly as the number of keypoints detected per second.

- *Semantic precision = number of keypoints detected on the image segment showing the subject / total number of keypoints detected both on image segment representing the subject and on the segment representing the background*. To get the first value, we apply the selected detection algorithm on the masked image. For the second value, the detector algorithm is applied to the original image, and the total number of detected features is calculated. This measure is used as an indicator of detector algorithm expressiveness – higher ratio indicates a better ability of the detector to distinguish between subject and background, and possibly reduce the number of bad matches and noise later in the reconstruction process.

4. Results and Discussion

This section offers our findings for detector evaluation on the human-based dataset.

4.1. Detector Evaluation

Here we present results of standalone detector evaluation for each of the aforementioned three metrics.

4.1.1. The average number of detected keypoints

As mentioned earlier, the number of detected keypoints can significantly impact later stages of the reconstruction process, since the low number of detected keypoints will undoubtedly lead to the low-quality avatar reconstruction. In Table 2, we show our findings on the average number of keypoints detected on the proposed dataset of seven scans for different scaling factors applied on both original and masked images. In all tested scenarios, binary detectors highly outperform SIFT and SURF, as shown in Table 3. In general, masking does not have a significant impact on the number of detected keypoints. We observe that the average number of keypoints detected on masked images can vary up to 10% compared to the average number of detected keypoints using the same detectors on original images. Since the change is positive in all cases except for SURF and MSD detectors, our estimate is that by eliminating the background from the input image, we emphasize contours of the subject which leads to the increased number of keypoints detected by the majority of detection algorithms. At the same time, keypoints detected on the background are discarded on masked images. In the case of SURF and MSD algorithms, the number of keypoints rejected by the mask is slightly larger than the number of newly detected keypoints on the masked images, which leads to the reduced number of keypoints detected on masked images. We observe that the average number of keypoints detected on masked images can vary up to 10% compared to the average number of detected keypoints using the same detectors on original images. In all cases except for SURF and MSD detectors, more keypoints are detected on the masked image than on the original image even though image masking discards all keypoints detected on the background. Higher keypoint detection rate on masked images can be contributed to additional keypoints being identified by the majority of detectors when subject contours enhancement is introduced.

Table 2 Detection algorithms ranked according to the average number of detected keypoints. In the first row are detectors that on average detect the largest number of keypoints, in the last row are detectors that on average detect the smallest number of keypoints. The table provides detector rankings on images with (column *Masked*) and without masks applied (column *Original*)

Rank	Scale 1		Scale 2		Scale 4		Scale 8	
	Original	Masked	Original	Masked	Original	Masked	Original	Masked
1	ORB	ORB	ORB	ORB	AGAST	AGAST	AGAST	AGAST
2	FAST	FAST	AGAST	AGAST	FAST	FAST	ORB	ORB
3	AGAST	AGAST	FAST	FAST	ORB	ORB	FAST	FAST
4	GFTT	GFTT	GFTT	GFTT	GFTT	GFTT	GFTT	GFTT
5	BRISK	BRISK	BRISK	BRISK	SURF	SURF	SURF	SURF
6	SIFT	SIFT	SURF	SURF	BRISK	BRISK	BRISK	BRISK
7	SURF	SURF	SIFT	SIFT	SIFT	SIFT	SIFT	SIFT
8	AKAZE	AKAZE	AKAZE	AKAZE	AKAZE	AKAZE	AKAZE	MSD
9	MSD	MSD	MSD	MSD	MSD	MSD	MSD	AKAZE
10	MSER	MSER	MSER	MSER	MSER	MSER	MSER	MSER
11	STAR	STAR	STAR	STAR	STAR	STAR	STAR	STAR

Table 3 Average number of detected keypoints

Rank	Detector	Scale 1	Scale 2	Scale 4	Scale 8
1	ORB	122591	46600	10392	2823
2	FAST	115551	39980	10570	2644
3	AGAST	114616	42207	11880	3013
4	SIFT	48546	9981	1872	608
5	SURF	37910	11416	3564	1079

4.1.2. The average number of detected keypoint per second

Since we do not limit the number of detected keypoints, it would be unfair to rank detectors directly according to the execution time. Instead, we use a relative ratio of the number of detected keypoints and time spent on keypoint detection, as shown in Table 4. Binary detectors FAST, AGAST and ORB show the best overall detection speed performance. Both on the original and masked images, FAST detects 3.5 and 4.5 times more keypoints per seconds then AGAST and ORB, respectively. Detected keypoint ratio between these three detectors persists even across different scales, which is not valid for the comparison of FAST and state-of-the-art SIFT and SURF detectors, where ratio variations are not negligible. For original images and different values of a scale factor, FAST detects up to 540 times more keypoints per second then SIFT, and 137 times more than SURF. For masked images, these ratios are slightly larger. When compared to other detectors, MSD and MSER algorithms are highly inefficient, detecting approximately less than a single keypoint per second on the original, and one to two keypoints per seconds on masked images.

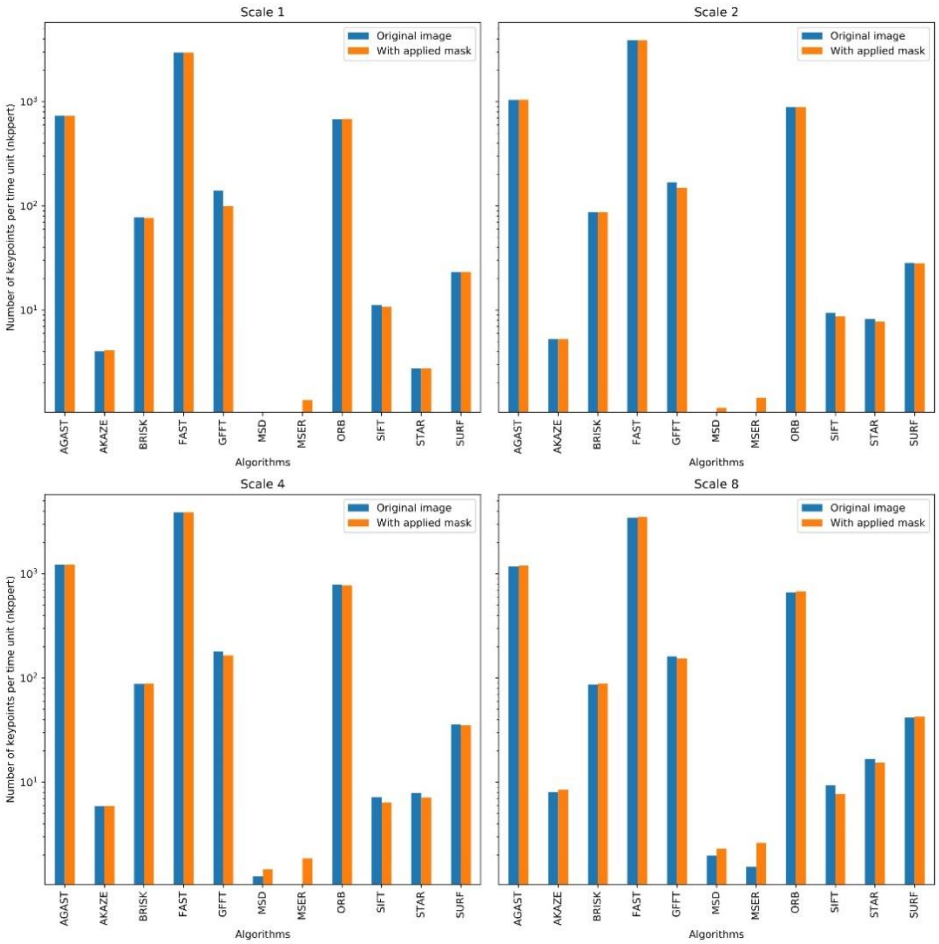


Fig. 3 Average number of keypoints detected per time unit (second)

Table 4 Average number of keypoints detected per second

Detector	Scale 1		Scale 2		Scale 4		Scale 8	
	Original	Masked	Original	Masked	Original	Masked	Original	Masked
AGAST	740	739	1046	1047	1224	1232	1172	1196
AKAZE	4	4	5	5	6	6	8	8
BRISK	78	77	87	87	88	88	86	88
FAST	2959	2956	3902	3878	3900	3904	3465	3516
GFFT	140	100	169	150	181	165	161	155
MSD	1	1	1	1	1	1	2	2
MSER	1	1	1	1	1	2	2	3
ORB	680	686	894	885	787	778	663	681
SIFT	11	11	9	9	7	6	9	8
STAR	3	3	8	8	8	7	17	15
SURF	3	3	8	8	8	15	17	15

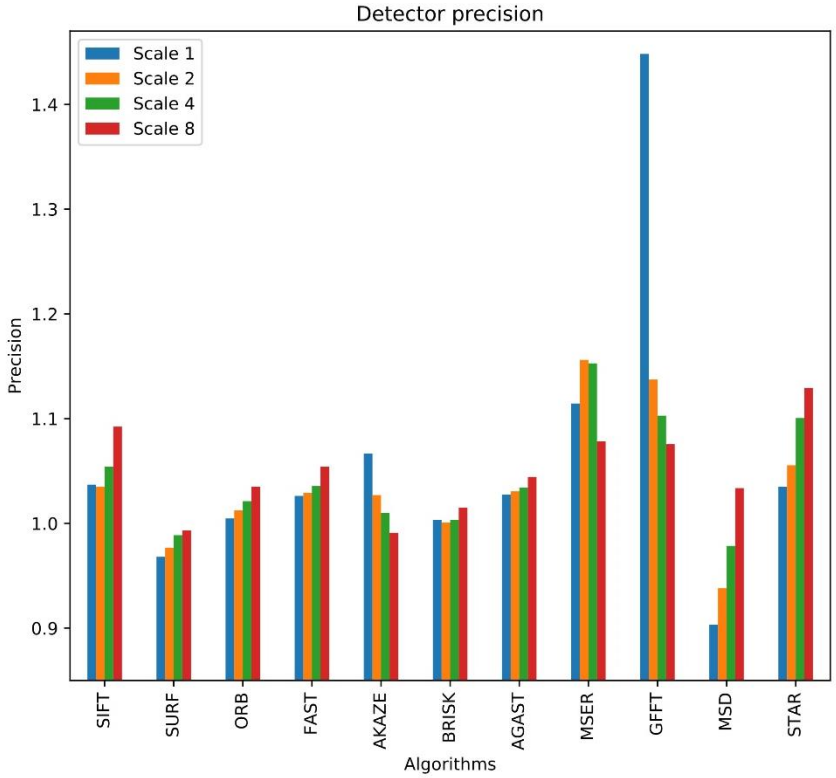


Fig. 4 The ratio of detected keypoints on original scans and scans with applied masks

4.1.3. Semantic precision

Not all keypoints are equally important in the process of human reconstruction since we would like to reconstruct the avatar of the subject with as little background noise as possible. That makes keypoints detected on the subject more important than the keypoints detected on the background. In Fig. 4 we show a ratio of the average number of keypoints detected on masked images and those detected on the original image. SURF and MSD are more likely to detect keypoints on the background for all tested values of the scale factor. Other algorithms express moderate to high robustness to the background keypoints since the computed ratio indicates that the number of keypoints detected on the subject is at least equal or even larger than the total number of keypoints detected on the original image.

5. CONCLUSION

In this paper, we presented an extensive evaluation of algorithms for keypoint detection in the context of 3D avatar reconstruction from an image sequence. Although similar exhaustive evaluations of detector performance exist, we are not aware of any other study performed in the context of photogrammetry-based human body reconstruction. First, we created a human body image dataset by capturing images of seven different persons in a multi-view stereo

setup in controlled lighting conditions. The dataset is used to evaluate eleven algorithms for keypoint detection, including well established and patented SIFT and SURF algorithms as a baseline. Our findings are mainly in agreement with previously conducted work proposed in [13], [17]. Binary detectors show superior performance compared to floating-point detectors in terms of detection speed and number of detected keypoints. Among the binary detectors, FAST is the most efficient in terms of speed detection, detecting a considerably larger number of keypoints per second comparing to SIFT and SURF detectors, followed by ORB and AGAST. ORB, AGAST, and FAST are top-performing detectors considering the number of detected keypoints; their performance additionally increased when performed on the masked image. In our use case, FAST does not produce the largest number of keypoints but is significantly close to the top-performing ORB detector with approximately 2% less keypoints detected. We also found that SURF and MSD in comparison with other detectors, discover a significant number of keypoints in the background area, meaning that the usage of this detectors in the pipeline could lead to noisy reconstructions.

In future work, detectors learned by machine learning techniques will be included in the evaluation. Although advanced handcrafted detector algorithms still exhibit at least comparable performance to those that are learned, machine learning is a rapidly developing area and it can be expected that learned detectors will outperform handcrafted soon. Another direction for future work includes improvement of the evaluation framework. The most reliable way to estimate actual detector performance would be to produce a 3D reconstruction based on detected keypoints. Current photogrammetry-based software commonly includes just SIFT and SURF detection algorithms into the pipeline. More work toward the adaption of other detectors in the pipeline will be done to additionally verify given numerical results.

Acknowledgements.*The research reported in this paper is partially supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia, projects number TR32044 (2011-2020), ON174026 (2011-2020), and III44006 (2011-2020).*

REFERENCES

[1] J. N. Bailenson, N. Yee, J. Blascovich, and R. E. Guadagno, "Transformed social interaction in mediated interpersonal communication", *Mediated Interpersonal Communication*, 2008, pp. 77–99.

[2] H. Lin and H. Wang, "Avatar creation in virtual worlds: Behaviors and motivations", *Comput. Human Behav.*, vol. 34, pp. 213–218, May 2014.

[3] F. Cordier, W. Lee, H. Seo, and N. Magnenat-Thalmann, "From 2D Photos of Yourself to Virtual Try-on Dress on the Web," In People and Computers XV—Interaction without Frontiers, London: Springer London, 2011, pp. 31–46.

[4] C. Zizza, A. Starr, D. Hudson, S. S. Nuguri, P. Calyam, and Z. He, "Towards a social virtual reality learning environment in high fidelity," In Proceedings of the 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2018, pp. 1–4.

[5] D. Dragan, Z. Anišić, S. Mihić, and V. Puhalac, "3D Avatar Platforms: Tomorrow's Gateways for Digitized Persons into Virtual Worlds", Springer, Cham, 2018, pp. 141–155.

[6] I. Hudson and J. Hurter, "Avatar types matter: Review of avatar literature for performance purposes," In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9740, pp. 14–21.

[7] M. Yuan, I. R. Khan, F. Farbiz, S. Yao, A. Niswar, and M.-H. Foo, "A Mixed Reality Virtual Clothes Try-On System", *IEEE Trans. Multimed.*, vol. 15, no. 8, pp. 1958–1968, Dec. 2013.

[8] T. Luhmann, S. Robson, S. Kyle, and J. Boehm, *Close Range Photogrammetry and 3D Imaging*. 2013.
[9] AgiSoft, "AgiSoft PhotoScan Professional (Version 1.2.6) (Software)", 2016. [Online]. Available: https://www.agisoft.com/downloads/installer/.
[10] J. Heinly, E. Dunn, and J. M. Frahm, "Comparative evaluation of binary features", In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, vol. 7573 LNCS, no. PART 2, pp. 759–773.
[11] AliceVision, "Meshroom: A 3D reconstruction software." 2018.
[12] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "Openmvg: Open multiple view geometry," In Proceedings of the International Workshop on Reproducible Research in Pattern Recognition, 2016, pp. 60–74.
[13] H. Aanæs, A. L. Dahl, and K. S. Pedersen, "Interesting interest points: A comparative study of interest point performance on a unique data set", *Int. J. Comput. Vis.*, vol. 97, no. 1, pp. 18–35, Mar. 2012.
[14] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.
[15] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors", *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, Oct. 2004.
[16] O. Miksik and K. Mikolajczyk, "Evaluation of Local Detectors and Descriptors for Fast Feature Matching," In Proceedings of the 21st Int. Conf. Pattern Recognit. (ICPR), 2012, Icpr, pp. 2681–2684, 2012.
[17] A. Canclini, M. Cesana, A. Redondi, M. Tagliasacchi, J. Ascenso, and R. Cilla, "Evaluation of low-complexity visual feature detectors and descriptors", In Proceedings of the 18th International Conference on Digital Signal Processing, DSP 2013, 2013, pp. 1–7.
[18] Ş. Işık, "A Comparative Evaluation of Well-known Feature Detectors and Descriptors," *Int. J. Appl. Math. Electron. Comput.*, vol. 3, no. 1, p. 1, Dec. 2014.
[19] D. Mukherjee, Q. M. Jonathan Wu, and G. Wang, "A comparative experimental study of image feature detectors and descriptors," *Mach. Vis. Appl.*, vol. 26, no. 4, pp. 443–466, May 2015.
[20] K. Yamada and A. Kimura, "A performance evaluation of keypoints detection methods SIFT and AKAZE for 3D reconstruction," In Proceedings of the 2018 International Workshop on Advanced Image Technology, IWAIT 2018, 2018, pp. 1–4.
[21] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes", *ACM Trans. Graph.*, vol. 22, no. 3, p. 587, 2003.
[22] A. S. Jackson, C. Manafas, and G. Tzimiropoulos, "3D Human Body Reconstruction from a Single Image via Volumetric Regression", Sep. 2018.
[23] G. Varol *et al.*, "BodyNet: Volumetric inference of 3D human body shapes," In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11211 LNCS, pp. 20–38.
[24] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D Human Reconstruction from a Single Image," Mar. 2019.
[25] A. Venkat, S. S. Jinka, and A. Sharma, "Deep Textured 3D Reconstruction of Human Bodies," Sep. 2018.
[26] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3D full human bodies using kinects", *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 4, pp. 643–650, Apr. 2012.
[27] Z. Liu *et al.*, "3D real human reconstruction via multiple low-cost depth cameras", *Signal Processing*, vol. 112, pp. 162–179, Jul. 2015.
[28] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun, "Multi-view image and ToF sensor fusion for dense 3D reconstruction", In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009, 2009, pp. 1542–1546.
[29] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[30] A. Weiss, D. Hirshberg, and M. J. Black, "Home 3D body scans from noisy image and range data," In Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1951–1958.
[31] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited", In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[32] J. L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo", In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9907 LNCS, pp. 501–518.
[33] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-Scale Data for Multiple-View Stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, Nov. 2016.
[34] M. Goesele, B. Curless, and S. M. Seitz, "Multi-View Stereo Revisited," In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06), vol. 2, pp. 2402–2409.
[35] S. R. Fanello *et al.*, "UltraStereo: Efficient learning-based matching for active stereo systems," In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017-Janua, pp. 6535–6544.
[36] I. Stančić, M. Brajović, I. Orović, and J. Musić, "Compressive sensing for reconstruction of 3D point clouds in smart systems," In Proceedings of the 24th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2016, 2016, pp. 1–5.
[37] V. Tan, I. Budvytis, and R. Cipolla, "Indirect deep structured learning for 3D human body shape and pose prediction," In Proceedings of the British Machine Vision Conference 2017, 2017.
[38] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-End Recovery of Human Shape and Pose," In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 7122–7131.
[39] D. Gajić, S. Mihić, D. Dragan, V. Petrović, and Z. Anišić, "Simulation of photogrammetry-based 3D data acquisition," *Int. J. Simul. Model.*, vol. 18, no. 1, 2019.
[40] G. Bradski, "The OpenCV Library," *Dr. Dobb's J. Softw. Tools*, 2000.
[41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[42] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features", In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006, vol. 3951 LNCS, pp. 404–417.
[43] M. Agrawal, K. Konolige, and M. R. Blas, "CenSurE: Center surround extremas for realtime feature detection and matchin", In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5305 LNCS, no. PART 4, pp. 102–115.
[44] F. Tombari and L. Di Stefano, "Interest points via maximal self-dissimilarities", In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9004, pp. 586–600.
[45] P. E. Forssén, "Maximally stable colour regions for recognition and matching", In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[46] D. Nistér and H. Stewénius, "Linear time maximally stable extremal regions," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5303 LNCS, no. PART 2, pp. 183–196.
[47] Jianbo Shi and Tomasi, "Good features to track", In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94, 1994, pp. 593–600.
[48] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF", In Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2564–2571.

[49] E. Rosten and T. Drummond, “Fusing points and lines for high performance tracking”, In Proceedings of the IEEE International Conference on Computer Vision, 2005, vol. II, pp. 1508–1515.

[50] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary Robust invariant scalable keypoints”, In Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2548–2555.

[51] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, “Adaptive and generic corner detection based on the accelerated segment test”, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6312 LNCS, no. PART 2, pp. 183–196.

[52] P. Alcantarilla, J. Nuevo, and A. Bartoli, “Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces”, In Proceedings of the British Machine Vision Conference 2013, 2014, pp. 13.1-13.11.