Exploring the Impact of Color Space in 6D Object Pose Estimation

Nuno Pereira
nuno.pereira@ubi.pt
Luís A. Alexandre
luis.alexandre@ubi.pt

Departamento de Informática
Universidade da Beira Interior
NOVA LINCS
6201-001, Covilhã, Portugal

Abstract

6D pose estimation is an open challenge due to complex world objects and many possible problems when capturing data from the real world, *e.g.*, occlusions, and truncations. Getting the best input data to the deep learning methods is critical, for example, light can alter the features that these methods extract from the objects. Not obtaining accurate poses of the objects can lead to poor experiences in augmented reality scenarios or can lead to a fail grasping task of a robot. To try to avoid these issues, we investigate the impact of color spaces in 6D object pose estimation. For that, we evaluated RGB, Grayscale, HSV, and the HSV individual channels to study which color space would perform better in the 6D pose estimation task. We increased the accuracy of a method in 7.11% by using the HSV color space instead of the frequently used RGB.

1 Introduction

In computer vision, the ambient light can be a notable problem. It can create artifacts, alter the colors or cause shadows in the captured scene therefore constituting a problem in many computer vision algorithms.

The RGB color space is widely used, although it does not represent the color as humans perceive it. If we want to isolate an object just using color in the image, it is hard to do in RGB because there may be many similar colors in the image.

The HSV color space has three channels similar to RGB but instead of Red, Green, and Blue we have Hue, Saturation and Value, or intensity. The Hue channel represents the color. For example, red is a color but light red or dark red is not. The saturation channel is the amount of color present. It differentiates the pale red from the pure red. Finally, the value or intensity represents the brightness of the color, light red or dark red. So in the Hue channel, each color has its own value the entire red is a particular value. The lightness or darkness of the color does not affect the hue channel, so this channel is useful to extract specific colors from images. In real photographs, you will obtain varied saturation throughout the images depending on the intensity of the color present in them. The intensity channel shows the brightness of the colors and this channel usually has much influence by the light source.

With the required automation needed to help humans in production lines, robots need to work in a collaborative mode and work in non-restricted environments so the capacity to understand the scene and the objects within is becoming a must. The most common task that robots do is object grasping, which is a task that has been tackled by many researchers because it needs to be as fast as possible and precise so there is no damage in the objects. Performing gasping in a non-restricted and cluttered environment, *e.g.*, bin picking, is a complex problem to tackle.

6D pose estimation is a task in computer vision that detects the 6D pose (3 degrees of freedom for the position and the other 3 for orientation) of an object. A 6D pose is as important in robotic tasks as in augmented reality, where the pose of real objects can affect the interpretation of the scene and the pose of virtual objects can also improve the augmented reality experience. It can also be useful in human-robot interaction tasks like learning from demonstration and human-robot collaboration.

Estimating the object's 6D pose is a challenging problem due to the diversity of objects that exist and how they appear in the real world. Obtaining the data to retrieve the 6D pose is a problem, as RGB-D data can be hard to obtain for certain types of object, *e.g.*, fully metallic objects, and meshed office garbage bins. Another problem during the data capture is the light present in the scene because it can generate noise or reflection in the objects. Distinct light sources can make deep learning methods extract different features from the same object this being a problem because if we want that the method learns these types of light sources we need to capture data from each type of light source. More specific, 6D pose estimation requires massive amounts of data to have good performance in real-world applications and even when the methods are trained in these big datasets they usually tend to fail or have a greater error in the real world because the most common and well-structured datasets have well-controlled light sources.

Figure 1: Qualitative results on the LineMOD Dataset. These results were obtained using MaskedFusion with HSV color space as input data. The red dots represent the object keypoints of the estimated 6D pose projected onto the RGB image.

To prevent this situation, we propose an alternative approach to this type of method by analyzing other color spaces. Color spaces like HSV are uncommon in the 6D pose estimation area of research. We test if using other color spaces in the most common dataset LineMOD [1] will increase the performance of 6D pose estimation methods while not increasing significantly its training or inference times.

2 Methodology

We use MaskedFusion [4] as a framework for 6D pose estimation in our experiments. It is one of the best-performing methods in the state-of-the-art.

MaskedFusion consists of three sub-tasks that executed sequentially estimates the 6D pose of an object presented in the scene. Initially, it uses a semantic segmentation method to detect and generate masks for each object presented in the scene. Then for each object segmented it crops the RGB image, depth image and mask. To eliminate the background around the object, a bit-wise and operation is made between the images and the mask. These segmented images are fed to a fully convolution neural network so it can regress the 6D pose of that object. After the preliminary pose is estimated, it is possible to utilize another method to refine the pose of the object. The method used in MaskedFusion is a neural network that enables it to be executed in real-time instead of other methods that are resource-heavy.

This work was supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT-Fundação para a Ciência e a Tecnologia, through national funds, and partially supported by project 026653 (POCI-01-0247-FEDER-026653) INDTECH 4.0 – New technologies for smart manufacturing, cofinanced by the Portugal 2020 Program (PT 2020), Compete 2020 Program and the European Union through the European Regional Development Fund (ERDF)

Table 1: Results presented in this table were obtained through the training of MaskedFusion with its weights initialized as random. Italic names represent the symmetric objects. Bold values are the higher values in each line.

Objects	RGB	Grayscale	HSV	H	S	V
ape	74.29	86.67	**97.14**	67.62	34.29	82.86
bench vi.	99.03	**100.00**	99.03	88.35	89.32	99.03
camera	96.08	**98.04**	**98.04**	87.25	75.49	97.06
can	94.06	97.03	**98.02**	80.20	91.09	93.07
cat	**97.00**	95.00	**97.00**	81.00	86.00	**97.00**
driller	96.00	95.00	**99.00**	91.00	88.00	94.00
duck	62.26	93.40	**96.23**	51.89	35.85	88.68
eggbox	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
glue	**100.00**	99.03	**100.00**	**100.00**	**100.00**	**100.00**
hole p.	91.43	95.24	**99.05**	89.52	74.29	**99.05**
iron	75.26	97.94	95.88	94.85	91.75	**98.97**
lamp	**100.00**	**100.00**	98.08	95.19	97.12	**100.00**
phone	**100.00**	99.04	**100.00**	93.27	94.23	99.04
Average	91.17	96.63	**98.28**	86.08	81.14	96.03

In our experiments, we did not use the first sub-task of the MaskedFusion. Our primary goal is to report the impact of the different color spaces and/or channels in the 6D pose estimation. Since MaskedFusion is a modular framework, it was effortless to remove the semantic segmentation sub-task and use the ground truth masks to make the operations for the crop and background removal.

To perform our tests, we choose to compare the HSV color space and each of its channels with the RGB color space. We tested MaskedFusion using the RGB, HSV, Grayscale, H (Hue), S (Saturation), and V (Value). We evaluated the MaskedFusion method in two independent roundups. In the first series of tests executed we trained the method from scratch, this means, the neural network presented in the method started with random weights, and we trained it for 150 epochs. In the second series of tests, we trained the method with RGB for 350 epochs. Furthermore, we saved the best performing weights in the validation set and use these weights to start fine-tuning the neural network for the other color channels. We fine-tuned the neural network for 150 epochs.

In our tests, we use the LineMOD [1] dataset because it is widely utilized in this area of research. It consists of 13 objects in over 18000 real images with the ground truth pose annotated. These images were captured with a Kinect camera that automatically aligns the RGB and depth images.

As in previous works in 6D pose estimation [2], [3], [5], [6], [4] we use the same evaluation metrics for the LineMOD dataset. The Average Distance of Model Points (ADD) [1] is used for non-symmetric objects and for the egg-box and glue the Average Closest Point Distance (ADD-S) [6] is used.

$$\text{ADD} = \frac{1}{m} \sum_{x \in M} \left\| (Rx + t) - (\hat{R}x + \hat{t}) \right\| \tag{1}$$

In the ADD metric (equation 1), assuming the ground truth rotation R and translation t and the estimated rotation $\tilde{R}$ and translation $\tilde{t}$, the average distance calculates the mean of the pairwise distances between the 3D model points of the ground truth pose and the estimated pose. In equation (1) and (2) M represents the set of 3D model points and m is the number of points.

For the symmetric objects (egg-box and glue), the matching between points is ambiguous for some poses. So for these cases, the ADD-S metric is used:

$$\text{ADD-S} = \frac{1}{m} \sum_{x_1 \in M} \min_{x_2 \in M} \left\| (Rx_1 + t) - (\hat{R}x_2 + \hat{t}) \right\| \tag{2}$$

3 Results

In Table 1 and 2, we present the results of MaskedFusion in the LineMOD test set. These results were calculated using the ADD (equation 1) and ADD-S (equation 2) metric. In Table 1, we present the results where the MaskedFusion neural network was trained for 150 epochs with weights initialized as random values.

In Table 1, its shown that the best performing color space is the HSV, as it performed higher on average. Specially for the first object, it achieved less error overall. HSV color space also achieved the best accuracy in 10 out of 13 objects.

Table 2: Results presented in this table were obtained by fine-tuning. Italic names represent the symmetric objects. Bold values are the higher values in each line.

Objects	RGB	Grayscale	HSV	H	S	V
ape	88.57	90.48	96.19	92.38	**97.14**	94.29
bench vi.	**99.03**	97.09	**99.03**	97.09	**99.03**	**99.03**
camera	**99.02**	**99.02**	**99.02**	98.04	95.10	**99.02**
can	96.04	**98.02**	97.03	**98.02**	93.07	96.04
cat	99.00	99.00	**100.00**	99.00	**100.00**	99.00
driller	96.00	95.00	92.00	**98.00**	**98.00**	94.00
duck	95.28	91.51	94.34	**96.23**	93.40	91.51
eggbox	99.06	**100.00**	**100.00**	99.06	**100.00**	**100.00**
glue	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
hole p.	99.05	**100.00**	**100.00**	**100.00**	98.10	98.10
iron	97.94	96.91	**100.00**	97.94	93.81	95.88
lamp	99.04	99.04	99.04	**100.00**	98.08	**100.00**
phone	94.23	94.23	97.12	**98.08**	97.12	94.23
Average	97.08	96.93	**97.98**	**97.98**	97.16	97.01

In Table 2, we present the results for the executed tests with fine-tuning. The results presented were obtained by using the best-performed weights in the evaluation set during 350 epochs and then we used these weights to fine-tune the MaskedFusion for the other color spaces. Fine-tuning took 150 epochs and then we evaluate the method in the test set. On average the HSV color space and the Hue color channel had the lowest average error in the LineMOD dataset. Both of these colors had seven objects in which they performed higher than the other color spaces/channels.

During inference, we took an average 0.014 seconds to estimate the 6D pose of an object. Our experiments took on average more 0.002 seconds to estimate the 6D pose of an object comparing its execution time with the RGB color space that did not need any color space conversion. These times were obtained using a computer with SSD NVME, 64GB of RAM, an NVIDIA GeForce GTX 1080 Ti and Intel Core i7-7700K CPU.

4 Conclusion

Sometimes using different color spaces aid in specific computer vision tasks. In these evaluations we discovered that using the HSV color space can help MaskedFusion achieve less error overall, if the same number of training epochs are used as when training using RGB images.

Training MaskedFusion for 150 epochs from the random weights in the RGB color space we achieved on average 91.17% accuracy and using the same setup but only changing to HSV color space we achieved 98.28%, a substantial improvement.

These tests might even achieve better results when dealing with real-world scenarios, since, the LineMOD dataset, as others, used a controlled light environment during the capture of the data creating the best possible scenario for each image presented in it. We suspect that the advantage of HSV over RGB can be even greater when pose estimation is performed in uncontrolled environments, and this will be a topic for future work.

References

[1] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*, 2011.

[2] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017.

[3] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

[4] Nuno Pereira and Luís A. Alexandre. MaskedFusion: Mask-based 6d object pose estimation. In *19th IEEE International Conference on Machine Learning and Applications (ICMLA 2020)*, December 2020.

[5] Chen Wang, Danfei Xu, Roberto Zhu, and Lu. Densefusion: 6d object pose estimation by iterative dense fusion. In *CVPR*, 2019.

[6] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv:1711.00199*, 2017.