ISSN 1870-4069

Depth Estimation Using Optical Flow and CNN for the NAO Robot

Oswualdo Alquisiris-Quecha[1], Jose Martinez-Carranza[1,2]

[1]Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Computer Science Department, Mexico
[2]University of Bristol, Computer Science Department, UK
oswualdoaq@inaoep.mx, carranza@inaoep.mx

Abstract. In the robotics field, one of the main challenges is autonomous navigation using a single camera where camera images are processed in a frame to frame basis. However, getting clear and noise-free images is still a challenge under erratic motion, typical of moving robots. To solve this problem, several works use optical flow techniques to eliminate blurred reference points in RGB images when the robot moves. The NAO robot is an example of a robotic platform that generates an oscillatory movement when walking, thus producing blurred images that may compromise the image processing task. In this work, we focus on the problem of depth estimation in a single image for the NAO robot, which proves useful for autonomous navigation. For the depth estimation, we argue that the erratic movement exhibited by the walking motion of the robot could be exploited to obtain optical flow vectors, which are strongly related to depth observed by the NAO's camera. Thus, we present a real-time system based on a Convolutional Neural Network (CNN) architecture that uses optical flow as input channels in order to estimate depth. To this aim, we present a new dataset that includes optical flow images associated to depth images for training. Our results indicate that optical flow can be exploited in humanoid robots such as NAO, but we are confident that it could be used in other platforms with erratic motion.

Keywords: depth estimation, deep learning, CNN, optical flow, NAO robot.

1 Introduction

One of the main challenges in the field of robotics is the autonomous navigation with a single camera, in which the images are processed frame by frame. However, due to the movement itself that produces a robotic system, such as the case of the NAO robot which generates an oscillatory movement when walking through the environment, it is not possible to obtain clear and noise-free images to be analyzed correctly in subsequent processes, therefore this remains a challenge under the erratic movements of moving robots.

To try to solve this problem, several techniques have been proposed, always trying to compensate the movement of the robot by optical and/or digital

stabilization on RGB images. However, when using this type of images it is possible to obtain blurred reference points when the robot moves, which is why optical flow techniques are used to solve this typical problem in RGB images

In this paper, we focus on the problem of depth estimation in a single image for the NAO robot, which is useful for the task of autonomous navigation. We argue that the erratic movement of a robotic system could be exploited to obtain optical flow vectors, which are strongly related to the depth observed by the camera. With this, it would no longer be necessary to use stabilization systems for the input images, freeing the computational resource that this implies for the robot, which is very useful for computational systems with little computing power like the case of the NAO robot.

Therefore, we present a system based on a CNN architecture that uses optical flow as input channels to estimate depth. For this purpose, we present a new dataset that includes optical flow images associated with depth images for training. Our results indicate that the optical flow can be exploited in humanoid robots like NAO, but we trust that it could be used in other platforms with erratic movement.

2 Related Work

Under the idea of autonomous navigation with the NAO robot, several investigations have been carried out trying to solve the problem of locating the robot within its environment to thereby achieve autonomous navigation. Initially, solutions were proposed by means of navigation based on estimates using environmental marks such as bar codes or landmarks [6,3,13] or using visual memories [1,4]. other works add additional sensors to the robot [15]. However, this considerably reduces the autonomy time of the robot and increases the instability of the system due to the added weight of the sensor—, which is not considered within the robot's kinematics. On the other hand, MonoSLAM based systems have been used for navigation where the robot incrementally constructs a map of an unknown environment in which it is located and in a parallel way estimates your trajectory of displacement in the environment through the use of a single camera as in [14,12].

However, a typical problem of these systems that work with RGB images is the obtaining of images with distortion and defocus due to the erratic movement of the robot when moving through the environment, for this reason, optical flow techniques have been used, which by its principle of operation works by having sequences of moving images.

This approach to employ optical flow techniques in conjunction with deep learning techniques, such as the case of Deep Learning, for the estimation of depth in monocular cameras, increasingly it is becoming a growing area to be used in mobile robot and humanoid as in [7] where they employ techniques that involve the analysis of the optical flow of a monocular image for the estimation of the depth and speed of displacement in an environment, similar to [11] where an architecture based on CNN is proposed for the estimation of distance of an

object in a 3D scene by using visual characteristics of optical flow,the network was trained with optical flow components, which has control signals for the evasion of obstacles in a mobile robot as system output.

3 Methodology

The estimation of depth in an environment using a single monocular image is an important task in autonomous navigation. Therefore, in this section, a general description of the proposed system for depth estimation is made using optical flow images for autonomous navigation tasks in humanoid robots such as the NAO robot. The general architecture is shown in Figure 1, in which the process of training and estimating the depth of the system is described in a general way.

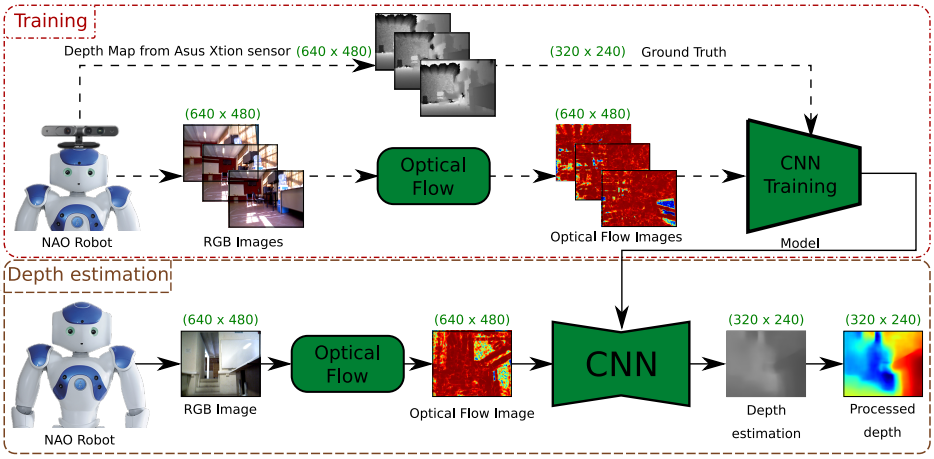

Fig. 1. General methodology.

The idea of using optical flow in depth estimation tasks is due to the fact that the flow vector of an image has a direct relationship with the relative depth of the objects in the image, besides that to obtain these values it is necessary to have sequences of images with movement. Movement that exists in every humanoid robot when moving around its environment, so it is proposed to use it in conjunction with deep learning techniques to generate depth maps of the environment.

The CNN network used is the DenseDepth proposed by [2] which is an encoder-decoder type network for depth estimation in RGB images, for the encoder part the RGB image is encoded in a feature vector using the DenseNet-169 and the decoder is composed of a successive series of ascending sampling layers with connections associated with the encoder without requiring any batch normalization, where the resolution of the input images of the network is 640 x 480 for the RGB and 320 x 240 for the depth maps estimated by it.

4 Description of the Dataset

In order to obtain images that correspond to those that the NAO robot perceives in its environment for the training of the neural network, it was necessary to search for the best alternative to gather the necessary data for the training phase. This is because the NAO robot only has one RGB camera and for the purposes of depth estimation it is necessary to generate a dataset where each RGB image has its corresponding depth map. Therefore, it was decided to use the Asus Xtion depth sensor, which was placed on the head of the NAO robot (See Figure 2) in order to obtain the images that the robot would see, also capturing the erratic movement of the device when moving in the environment.

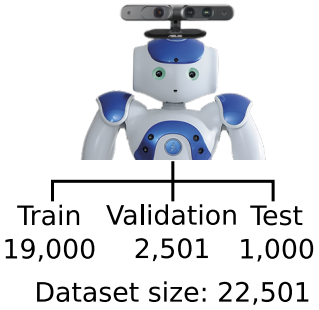

Fig. 2. Dataset from NAO image.

For the generation of the new dataset an implementation was made in C++ language which operates on the ROS system, where to access the sensor data the OpenNI software is used, considering a data synchronization policy of both RGB images and their corresponding depth images.

For RGB images, the developed system takes the I_{i-1} and I_i images from a sequence of images and the Farneback method is applied to calculate the optical flow for each pixel of the image, that is, a flow calculation is performed dense optical on the pair of input images obtaining an output image with the apparent movements of the objects that are inside the scene.

On the other hand, for the depth images corresponding to each calculated optical flow image, a process of removing outliers (negative values or NaN) is performed by assigning the value 0 in each pixel where there is an atypical value. The missing depth values are filled using the painting method proposed by [10], then a process of normalization of the data is carried out at a range of $[0 - 255]$ where the maximum distance corresponding to the depth is set to the 5.5 meters, the above through the equation (3):

$$I_i = (255 * D_i)/5.5, \tag{1}$$

where: I_i is the resulting depth image when applying normalization and D_i is the depth image without outliers.

The dataset is made up of 22,501 elements, with images obtained in interior scenes, of which they are divided into three groups: Training, Validation and Test, with 19,000, 2,501 and 1,000 data respectively (See Figure 2). Each group consists of RGB images, optical flow and depth map, each with a resolution of 640 x 480, which are each acquired at the same time. An example can be seen in Figure 3.

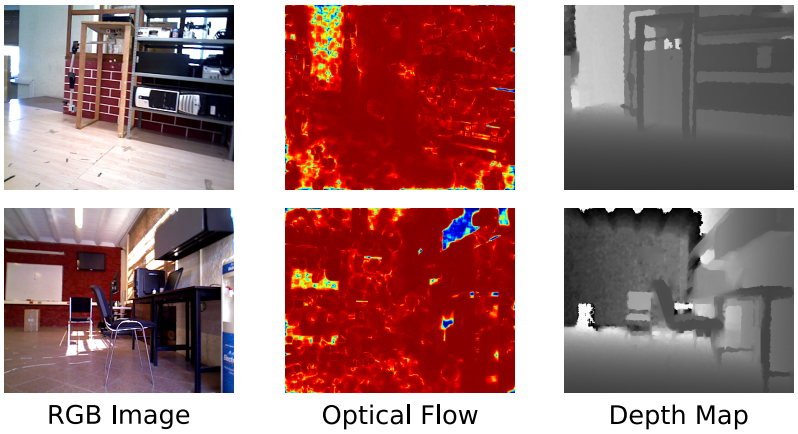

Fig. 3. Structure of the dataset.

5 Training

For the training of the CNN network, the dataset proposed in Section 4 is used, which contains images with interior scenes captured by the Asus Xtion sensor, which was placed on the head of the NAO robot to capture the images that the robot would see , also containing the erratic movement of the robot when navigating its environment. This dataset is composed of pairs of optical flow images and their corresponding depth image, in addition to the technique of data augmentation through transformations of training data whose technique has been shown to generate higher performance and obtain better accuracy. To do this, transformations of horizontal rotations are made to the images with a probability of 0.5, this is because vertical rotations in an image with an interior scene may not provide much information to the learning process and in some cases it could confuse the system because of the similarities of the geometries of floors and ceilings. In the same way different permutations are applied in the color channels with a probability of 0.25, this is considered according to the results obtained by [2] using this parameter.

For the training process, the CNN network is configured using the ADAM optimizer [8], which combines the methodology of Momentum and Root Mean Square Propagation (RMSProp), calculating a linear combination between the gradient and the previous increase, and considers the gradients recently appeared in the updates to maintain different learning rates per variable, with a learning rate value of 0.0001 and values $\beta_1 = 0.9$, $\beta_2 = 0.999$ (default values, optimizer's own), with a batch size of 1 for 50 epochs.

In addition, the loss function proposed by [2] is used, where this function seeks to balance the reconstruction of depth images by minimizing the difference of the ground truth values and at the same time penalizes the high frequency distortions in the image domain of depth. The loss function consists of three terms and is defined as:

$$L(y, \hat{y}) = \lambda L_{depth}(y, \hat{y}) + L_{grad}(y, \hat{y}) + L_{SSIM}(y, \hat{y}), \tag{2}$$

where:

$$\lambda = 0.1, \tag{3}$$

$$L_{depth}(y, \hat{y}) = \frac{1}{n} \sum_{p}^{n} |y_p, \hat{y_p}|, \tag{4}$$

$$L_{grad}(y, \hat{y}) = \frac{1}{n} \sum_{p}^{n} |g_x(y_p, \hat{y_p})| + |g_y(y_p, \hat{y_p})|, \tag{5}$$

$$L_{SSIM}(y_p, \hat{y_p}) = \frac{1 - SSIM(y, \hat{y})}{2}. \tag{6}$$

In the equation (5), g_x y g_y represent the differences in the x and y components for the gradients of the depth image of y and $\hat{y}$

6 Experiments and Results

The network implemented in TensorFlow for depth estimation using optical flow images of the scene was trained using a GeForce GTX 1050 GPU with 640 Cuda Cores, with the pre-trained weights of DenseDepth [2], using the learning transfer technique using the new dataset proposed in this paper.

6.1 Evaluation

For the quantitative evaluation the method is compared with other works of depth estimation that use RGB images as input, the evaluation is done using six evaluation metrics proposed by the state of the art. Where the error functions are defined as:

– Average Relative Error (REL): $\frac{1}{n} \sum_{p}^{n} \frac{|y_p - \hat{y_p}|}{y}$

- Root Mean Squared Error (RMS): $\sqrt{\frac{1}{n}\sum_p^n (y_p - \hat{y_p})^2}$
- Average (Log_{10}) error: $\frac{1}{n}\sum_p^n |log_{10}(y_p) - log_{10}(\hat{y_p})|$
- Threshold accuracy (δ_i): $\delta < thrforthr = 1.25, 1.25^2, 1.25^3$

Where y_p is the value of a pixel of the depth image y, $\hat{y_p}$ is the value of a pixel of the predicted depth image $\hat{y}$ by the trained model, and n is the total number of pixels for each depth image.

Table 1 shows this comparison, where it is necessary to denote that these models with which our proposal is compared are models obtained through various training processes using large RGB image data sets, in the order of millions of data. Our method and model obtained was trained only using 22.5k optical flow images using 50 training times through a training configuration from scratch, the values obtained can be low compared to the other methods with which it is compared, however it is necessary to consider mentioned above.

Table 1. Comparison with other methods of depth estimation.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL $\downarrow$	RMS $\downarrow$	$Log_{10} \downarrow$
Eigen et al. [5]	0.769	0.950	0.988	0.158	0.641	–
Laina et al. [9]	0.811	0.953	0.988	0.127	0.573	0.055
Alhashim et al. [2]	0.846	0.974	0.994	0.123	0.465	0.053
Ours	0.401	0.600	0.760	0.490	1.080	0.170

In Figure 4, the qualitative results of the system are shown. Where, the first column represents the RGB image only for questions of visual comprehension of the reader, the second column corresponds to the optical flow calculated for the RGB image which is the input to the CNN network, in the third column the Ground Truth is shown, the fourth column corresponds to the output of the CNN network and finally the last column represents the output image of the network by applying a representation in 3D color space for the image for better visual understanding.

Figure 5 shows a fragment of the navigation process of the NAO robot using the proposed system. In it, it is possible to observe the depth estimation obtained by means of optical flow images as input to the CNN network at moments of time during the navigation of the robot.

7 Conclusions and Future Work

In this work we propose the use of optical flow images as input to a CNN network for depth estimation, the idea of using optical flow is due to the fact that by its nature it can represent relative depth according to the values of the vector of optical flow and, in conjunction with a CNN network, we show that it is possible to estimate depth with metric of an arbitrary scene.

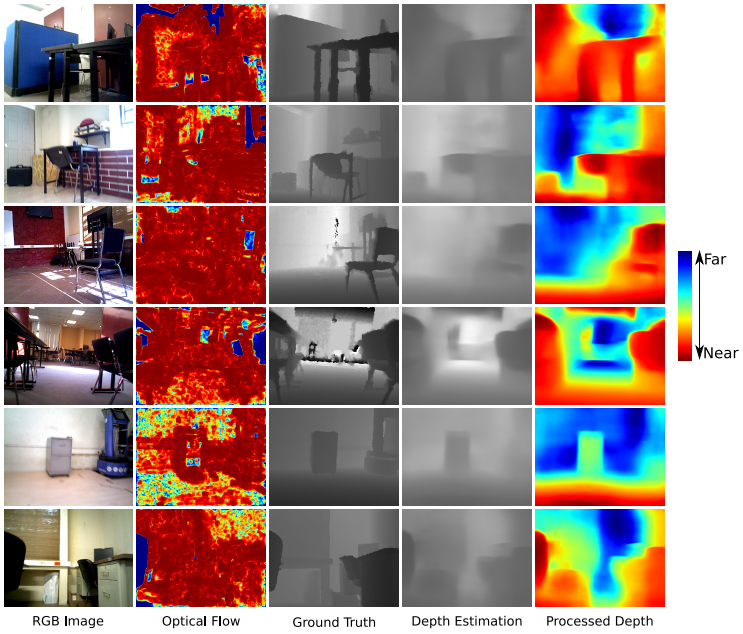

Fig. 4. Qualitative results from proposed method.

According to the experiments carried out, it is concluded that it is possible to use optical flow techniques in conjunction with deep learning techniques, such as the case of Deep Learning, for the estimation of depth maps in autonomous navigation tasks, where it is shown that it is possible to exploit the erratic movement in a humanoid robot to generate depth maps of its environment through the optical flow vectors generated by this movement.

In addition, a new dataset is proposed with images obtained from one of the NAO's frontal camera. In this dataset, we have collected color images obtained from mapping optical flow to the RGB space. The optical flow is generated during the walking motion of the NAO robot. These images are associated to corresponding depth images recorded with a Kinetc sensor. Thus, our main argument is that the erratic motion induced on the NAO's head while walking can be exploited by means of observing the optical flow, and such instantaneous flow, obtained in a frame-to-frame basis, can be exploited to learn depth relative to the robot. For the learning, we used a state of the art CNN architecture used for the problem of depth estimation in a single image, except that instead of using conventional RGB images, we propose to use our coded RGB images mapped from the optical flow. Our results indicate that our approach is feasible and it compares to state of the art methods on depth estimation in a single image.

As future work, we will explore new architectures of CNN networks in order to reduce processing time and to obtain better results in the estimated depth maps. Likewise, we will explore new data augmentation policies and probability

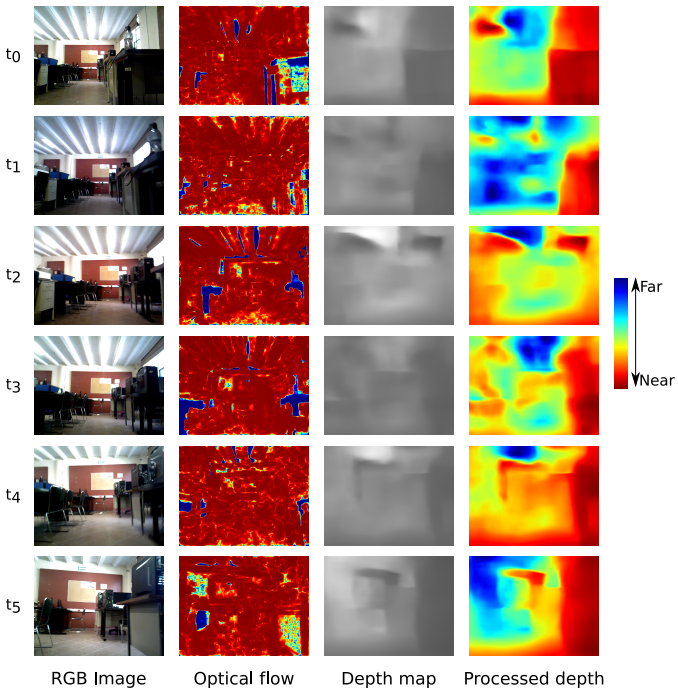

Fig. 5. Results of our proposed method obtained for a sequence of images during navigation of the NAO robot.

values that can help in the training process, generating a faster convergence and requiring a smaller number of iterations in the training process.

References

1. Aldana Murillo, N.G.: Localización de robots humanoides basada en apariencia a partir de una memoria visual. Tesis de Maestría en Optomecatrónica. Centro de Investigaciones en Óptica, A.C. León, Guanajuato p. 75 (2014)
2. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941 (2018)
3. Changyun, W., Junchao, X., Chang, W., Wiggers, P., Hindriks, K.: An approach to navigation for the humanoid robot nao in domestic environments. In: Towards Autonomous Robotic Systems: 14th Annual Conference, TAROS 2013, Oxford, UK, August 28–30, 2013, Revised Selected Papers. vol. 8069, p. 298. Springer (2014)
4. Delfin, J., Becerra, H.M., Arechavaleta, G.: Humanoid navigation using a visual memory with obstacle avoidance. Robotics and Autonomous Systems (2018)
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)

6. George, L., Mazel, A.: Humanoid robot indoor navigation based on 2d bar codes: Application to the nao robot. In: Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on. pp. 329–335. IEEE (2013)
7. Ho, H.W., de Croon, G.C., Chu, Q.: Distance and velocity estimation using optical flow from a monocular camera. International Journal of Micro Air Vehicles 9(3), 198–208 (2017)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
10. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM transactions on graphics (tog). vol. 23, pp. 689–694. ACM (2004)
11. Ponce, H., Brieva, J., Moya-Albor, E.: Distance estimation using a bio-inspired optical flow strategy applied to neuro-robotics. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2018)
12. Rioux, A., Suleiman, W.: Autonomous slam based humanoid navigation in a cluttered environment while transporting a heavy load. Robotics and Autonomous Systems 99, 50–62 (2018)
13. Wen, S., Zhang, Z., Ma, C., Wang, Y., Wang, H.: An extended kalman filter-simultaneous localization and mapping method with harris-scale-invariant feature transform feature recognition and laser mapping for humanoid robot navigation in unknown environment. International Journal of Advanced Robotic Systems 14(6), 1729881417744747 (2017)
14. Wirbel, E., Bonnabel, S., de La Fortelle, A., Moutarde, F.: Humanoid robot navigation: getting localization information from vision. Journal of Intelligent Systems 23(2), 113–132 (2014)
15. Xu, X., Hong, B., Guan, Y.: Humanoid robot localization based on hybrid map. In: Security, Pattern Analysis, and Cybernetics (SPAC), 2017 International Conference on. pp. 509–514. IEEE (2017)