Vision-Based Estimation of Altitude from Aerial Images

Rayan A. A. Alharazi , Muhammad A. Ahmad, Zaw Zaw Htike, Wai Yan Nyein Naing
Department of Mechatronics Engineering
Faculty of Engineering, International Islamic University Malaysia

ABSTRACT

One of the wide engineering fields is aircraft technologies and one of the most common needs for Airplane or UAV is estimating the altitude, which is some time difficult to estimate due to weather fluctuations and instability of the main parameters like pressure and speed. However, a combination of different sensors has been used to estimate altitude to guarantee an accurate reading and it is the method used these days. To overcome this problem is to use more capable technology such as machine vision based system to estimate the altitude, as advantages light weight, intelligence and accuracy, cheaper than commercial sensors as well as, computationally inexpensive. In this paper, we propose a vision-based system that can perform altitude estimation from aerial images. The satisfactory experimental results demonstrate the effectiveness of the proposed system.

Keywords: Altitude estimation, aerial images.

1. INTRODUCTION

Nowadays engineers tend to use modern technology to enhance the efficiency of their tasks in different fields. Machines and robots are built to make things easier and reduce the labor cost. As for human being, faulty and fatigue are common as it is a must to happened. One of the wide engineering fields is aircraft technologies and one of the most common needs for Airplane or UAV is estimating the altitude, which is some time difficult to estimate due to weather fluctuations and instability of the main parameters like pressure and speed. However, a combination of different sensors has been used to estimate altitude to guarantee an accurate reading and it is the method used these days. To overcome this problem is to use more capable technology such as machine vision based system to estimate the altitude, as advantages light weight, intelligence and accuracy, cheaper than commercial sensors as well as, computationally inexpensive.

We can notice the capability of this technology once we understand how a wild bird can fly, scan and hunt with main depending on the vision analyze and process. They didn't have any ultrasonic waves to send out or receive, they just use views from their eyes to estimate the distance of the prey, regardless other birds whose hunt at night like bats, etc.. Therefore, many research held to find an algorithm to build a depth map from a single image like what Saxena Et. Al. state in [1]. This study is examined to estimate the altitude from aerial image.

Let's us analyze the way we can estimate at first place. The images produced by the digital camera are referred as matrix of numbers; the image shows a simple 2D representation of the structures. The images produced from the camera hold very large and important information; such information can be very helpful for performing analysis, extracting features and result with an estimation.

But then this can't be done by the system itself. As in a picture, the image to be measured might be different in lighting, contrast, size, and surface. So with the same height but in different places, the system might estimate different altitudes. So for this problem, we might need to teach the system, using a supervised machine learning or maybe semi-supervised learning, so that the system will learn and store the data of each image and do a cross test to estimate the height of a different picture, places, surfaces and heights. With this, the estimation might be more accurate. Another idea is to learn the texture analysis and mapping of an image to extract information contained in the aerial picture and estimate the height value stated by the taken image. It may not be so accurate at first, but it will be when the system do have a lot of templates and can do a cross test for the images. The difficulty in this research is at first to train the system for lots of aerial images and their depths. Then when complying this with the real time aerial images, the quality of the images given might give some disadvantages and errors to the estimation. So an excessively many images with different contrast, quality, heights, and mostly different surfaces need to be taught to the system so that it can match the real time images given to some of the images in the database.

The estimation of altitude from an aerial image needs set of algorithm to enable the system to work properly. Therefore, to choose the proper approach we have to consider the following: (i) images are Top-Down aerial view. (ii) In aerial image we have to consider smaller structure and little details compared to other image processing taken in the ground. (iii) Image data

would contain outliers as there are irrelevant structure in shape and height on the snapped location, and we need to compute a single and major altitude for the aerial image [2].

In this paper we will search and study relevant paper and extract the best approach to solve our course project. High performance and accuracy are very important factors in estimating the altitude of aerial image, for that we must use the best and most appropriate tools for the job.

2. LITERATURE REVIEW

There are plenty of methods can be used to detect objects. However, using an effective method is the most important area of research in computer vision or image processing due to many applications needed, one paper proposed interesting approach which is "multi view Geometry based approach" it helps to construct a digital mapping form the aerial image taken for the ground. Let's investigate what have been researched so far related to the topic in recent years.

Vision based altitude estimation of an image has been done and investigated quite many in this era. Different camera arrangements and systems have been used to test the accuracy of the systems. A camera use normal lens for top-down view has been interduced in [4] and [11]. But then their altitude estimation approach is relative to a certain surface such as the landing base. On other hand paper [9], they suggested the geometry based approached to construct a map for the surface ground. But in a real time, surfaces and grounds are random and infinity, there might be trees, rocks, buildings, and millions of objects that will make the system unable to estimate an accurate altitude of the given image. One of the solutions to make a threshold for all unwanted detection and as assumption they consider the surface of the ground is flat. Therefore, the main task for the system is to result with only one altitude status from the entire aerial picture. In Fig-1 it shows the threshold for trees as they assume that trees look like a circular from top view.

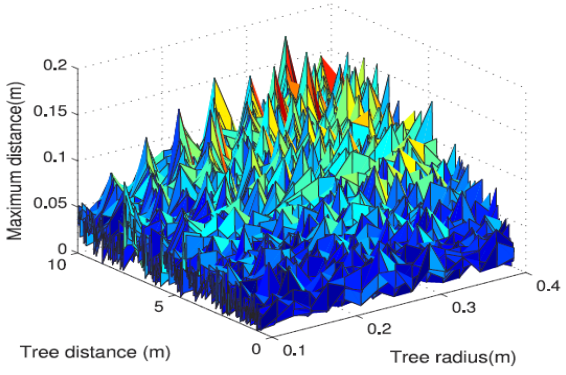


Fig-1. Segmentation threshold determination [13]

To remind how it is important to estimate the altitude, most of the airplanes around the world use a multi detection method to estimate its speed, altitude and attitude.

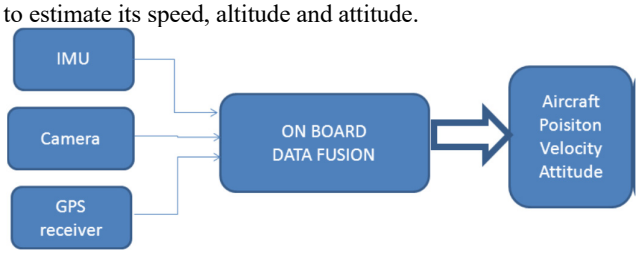


Fig-2. Data fusion architecture [3].

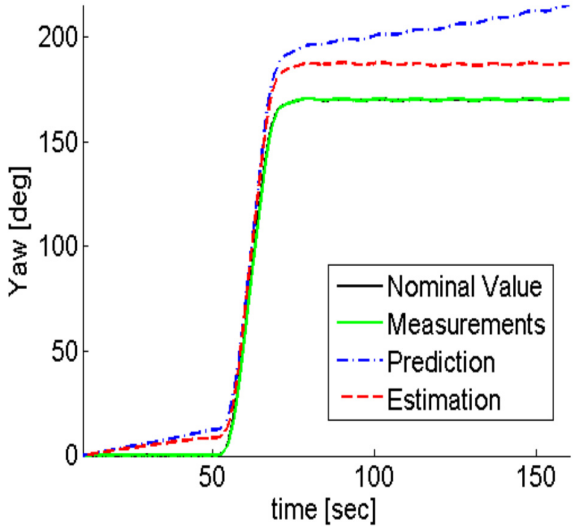
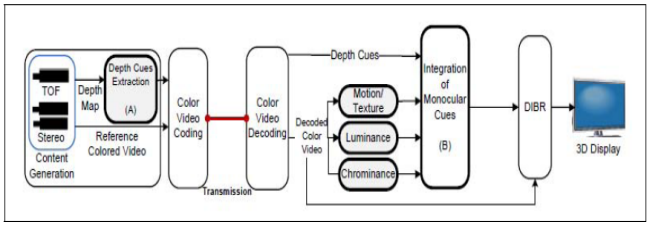


Fig-3. Yaw angle estimation, Nominal1 value (solid black line), Measurements1 (solid green line), 1Prediction1 (dashed blu line), 1Estimation (dashed red line)[3].

Fig-4. A block diagram demonstrating the proposed framework for depth maps extraction. The proposed algorithm has two components labeled as (A) and (B), respectively [16]



The use of inertial sensors in machine vision applications has been proposed, by now more than twenty years ago and further, studies have investigated the cooperation between inertial and visual, systems in autonomous, navigation of mobile robots, in image correction and to improve motion estimation for 3D reconstruction of structures. Recently, a framework for cooperation between vision sensors and inertial sensors has been proposed. The use of gravity as a vertical reference allows the calibration of focal length camera with a single vanishing point the segmentation of the vertical and the horizontal plane. In [4] is presented with a function for detecting the vertical (gravity) and 3D mapping, and in [5] such vertical inertial reference is used to, improve the alignment and registration of the depth, map.

For our project we are interested in camera operation and we would like to mainly depend on machine vision and how we could build a 3D mapping from an image, to the best of our knowledge, in paper [2] they deliberated entirely an elevation approximation for UAV and machine learning framework being advised. As motivation, recently paper [16] introduces 3D rebuilding using monocular cues see figure 4.Moreover, paper [6] and [7] again Saxena, Et. Al. he recommends a certain process for mapping the depth from only one picture. The procedure uses a "Markov Random Field" (MRF) based supervised learning to extract height variation model for each pixel in the aerial picture alongside with feature vectors collection analyzed from these pixels. However, their method cannot be applied to our task due to the following reasons: (i) aerial image has top-down view, (ii) structure in the aerial image are smaller related to normally taken on the ground pictures (iii) for simplification propose an assumption has been taken state that the earth surface is uniform, thus we need to calculate single elevation value of the entire aerial picture.

Moreover, in paper [2] they state that UAV does not undertake rapid variations in elevation so it confirms that the deviation of elevation from current picture to its previous patch pictures is small. They integrate these facts also into their model to produce a fine reading for altitude. They recommend a semi-supervised learning technique, from a quantity of possible landscape aerial pictures. This is similar to "Self-Taught learning" approach suggested in [8]. "Self-taught learning" is similar to transfer learning, this technique depend on analyzing the ingredients contained in the aerial picture like edges, textures, etc. And therefore, teaching the system by train and test random set of aerial pictures deliver linear combination of these bases which is a powerful representation model for any given picture. But they state that this method is not coefficient enough to give a proper result for UAV or aerial images. Eventually, they purposed a supervised learning method, a regression approach used to extract feature form aerial image and group it in hyper plan using the given set of elevation information.

Moreover, paper [17] they introduces "novel spatio-temporal MRF model" for altitude evaluation from set of pictures compared to the elevation of other patches in the same set of pictures and areas across this set of pictures in frames taken in earlier time. Later, the MRF model is resolved for the "Maximum A Posteriori" (MAP) approximation of the elevation. However, they begin to select a date like video altitude variations taken with a mobile device with a "fixed focal length", the researcher will not have much difficulty to infer the elevations between frames. For example, they believe they are able to easily tell if a photo was taken very close view to the earth surface or faraway, or how is the transformation in altitude between two image data. This is because of lack of prior knowledge of the environment, but also the possibility of using the tracks as monocular, changes in texture, the size of known objects, fog, emphasis / focus, etc. in inference. Distribution gradients texture capturing direction, the edges. It is a valuable source of depth cues and has been used very effectively in papers such as [6], [1] for the 3D reconstruction. When working with aerial images taken from a UAV, they exposed to more questions that cannot be sufficiently addressed by changes in texture only. For instance, most of the pictures are noisy, have a number of differences in lighting, or often are blurred by the movement of the UAV. Also, aerial picture got less structure compared to picture taken on the ground. For example, images of the earth, we may assume that there is a ground surface, all objects acquire a reference, etc. but the aerial images, views up and down like random dots and application of conventional filters such as filters auto - correlation filters, Fourier / wavelet based texture, slope, such as filters Nevatia - Babu, Lois masks, filters do not have enough capacity to efficiently capture, texture changes, the respective altitude changes. Figure.3. shows some sample images used in the experiment and Figure.4. Other images of the sample work on our project. They were taken to our campus and the height at which each image was taken is also stated. Note the change in texture with increasing altitude.

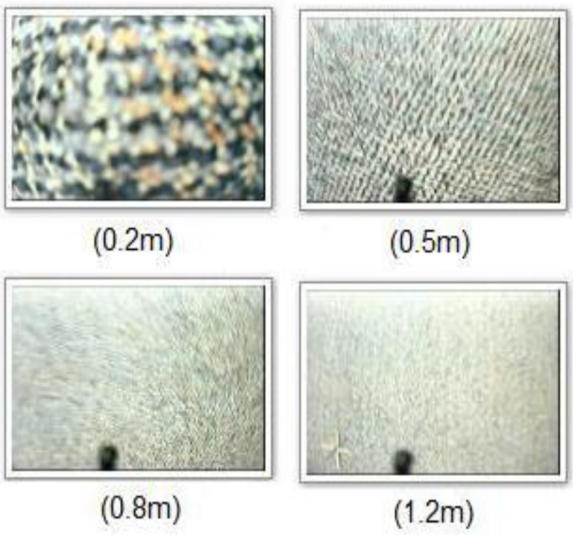


Fig-5. sample images collected from laboratory of the experiment for research paper [2]

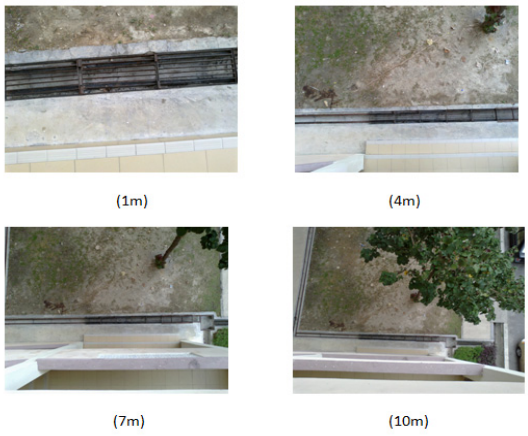


Fig-6. sample images collected from our university campus and hostel for our project.

From past research [12], they suggested a sparse over complete basis for an efficient framework for learning. This later used to classify objects. Our objectives are quite the same as them which we learn the basis from random images took from various places. But then the difference is that we took from aerial images, not random from any random angel. Thus our aim and philosophy is closer to the research in [17] where they use a semi-supervised learning. In order to approach our methods, we used a lot of aerial images from various places and types of surfaces to teach the system and build our basis set to train the system. Set of pictures are shown in Fig.4. above. There is also another research [4] that quite similar to our research where they use aerial images from a various fields of the Internet as the base assembly for the system. Some examples of images used in their research and for the same purpose as ours are shown in the figure.5.

Fig-7. Random aerial images from various places and terrains

To extract the features from images we need to apply some mathematics, regarding paper [2], they consider a large set of corpus of image patches I={I1,....,In}. Then it is vectorized to an approximately linear combination weight of basis vectors of n.

$$y^t \approx \sum_{j=1}^{n} b_j a_j^t = Ba^t \qquad (1)$$

(b1,...,bn ϵ Rk) are the basis vector and (ai ϵ Rn) is coefficient of vector sparse. Moreover, researches in this paper [2], they are using about 250 random topdown aerial images of the same size (640 x 480) for the machine learning part. Each of the images are first converted to a gray scale which is patched into size of 10 x 10. The images are then vectorized before they are used to build the basis by their formula (2). The equation used is very complex and yet to be studied if it is needed to be used in our system. As far as we know, y^i is the input, b_j is a basis linear weight combination, a^i_j is the corresponding weights activation. Fig.-8. shows the average absolute error in the altitude estimation versus basis set size plots.

$$\min_{b,a} \sum_i \|y^i - \sum_j a_j^i b_j\|_2^2 + \beta\|a^i\|_1 \qquad (2)$$

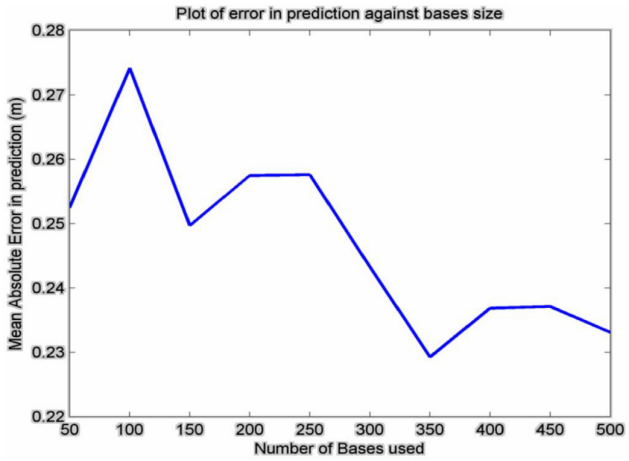


Fig-8. Mean absolute error of estimation versus number of bases used plot

The objective of optimization in (2) gives two different terms. (i) for each input yi to be well reconstructed, it encourages the first quadratic term. Then the basis's weighted linear combination act as bj and the corresponding weights given by the activations act as aij. (ii) the activation is encouraged to have low L1, so that ai become sparse. They state that the problem about the optimization is they convex over the variables a and b. But then the convex is not jointly to each other. For extra details, constrained least squares problem for L1 which activate a, on other hand, regularized least square problem for L2 on the basis b. In paper [9] they provide the algorithms to solve these 2 problems already. Fig. 3 illustrate how these algorithm effect in the basis set by using random topdown aerial images captured.

As they obtain the sparse basis set B ϵ Rk × n, now they state "the feature vector, f can be constructed for image dimension, k of the image patch, p". And this is the equation,

$$\min_f \|p - \sum_j f_j b_j\|_2^2 + \beta\|f\|_1 \quad (3)$$

The image Fig.9. below shows the stacked up of all the vectors' features, f from the given image and forming the feature vector set F.

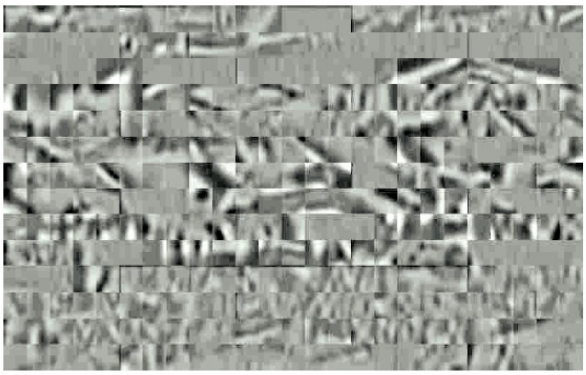


Fig-9. 350 basis vectors from random aerial images taken from the internet.

Now after some basis linear combination of the texture is represented as an image, the Gaussian Markov Random Field (MRF) is modeled by a supervised learning [10]. The MRF is used to estimate the every image's pixel block posterior altitude distribution [10]. The altitude posterior distribution, d is modeled given the vector set of the feature, F, and the σ and θ parameters as:

$$P(d|F;\sigma,\theta) = \frac{1}{Z} exp(-E_{\sigma,\theta}(d,F)) \quad (4)$$

where

$$E_{\sigma,\theta}(d,F) = \sum_{i=1}^{n} \frac{(d^i - F_i'\theta)^2}{\sigma_a^2} + \sum_{j=1}^{T}\sum_{i=1}^{n} \frac{(d_{t-j}^i - d^i)^2}{\sigma_j^2} + \beta \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} |d^i - d^j| \quad (5)$$

They assume that there is no sudden change in altitude of the UAV. Thus there will be less drastic deviations in the captured image, and predicted altitude would be more accurate. This is formulated in (5) where it constrains the pixel block's altitude, i at t time, to be smooth. Then at time t – j, then pixel block is framed to the image, I. We are constraining the predicted altitudes from many pixel blocks, and assuming it is a flat ground with one single altitude value.

Then, say dj and di represents two pixel blocks altitudes in the same frame, thus (dj - di) would be very near possible to zero. Fig-10. show MRF setup scheme and we assume in our computations.

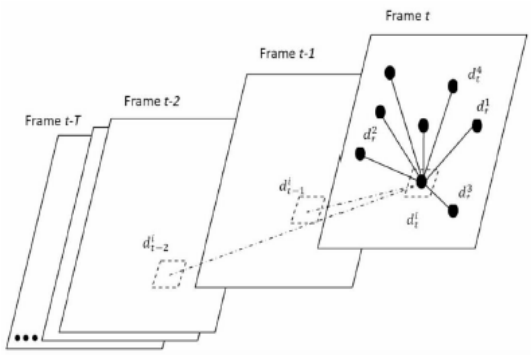


Fig-10. The MRF setup model taken from [10]. It represents the altitude dependency at the block of the pixel i to other blocks of the pixel. The frames are shown at times t, where t - 1, until t - T.

In [11], they address the depth estimation by presenting a stereo model that uses the constraints from some point with a known depth. With that they are referring to the Ground Control Points (GCPs). The formulation models are also influenced by the Markov Random Field. Different from our research we have no known depths as we are aiming to shoot any images from a random place and ground and let the system to estimate the altitude of the given aerial images. The methodology addressed in their research [11] is called the Stereo Matching. They first assume the scene rigidity and with known camera geometry, then the stereo matching algorithms will estimate a three-dimensional scene structure with several images taken from some different viewpoints. The common assumption used in the images is Lambertian or brightness constancy assumption where the images will appear equally in brightness aspects from any direction.

There is this one past research on height estimation using an aerial side images sequence [12] consider a "spotlight sequence" configuration view to estimate the height of the aerial images. The "spotlight sequence' is the control of the camera orientation throughout the acquisition. It is to maximize the image recovering part by maintaining the scene`s particular point by the center of the field. This configuration is as in Fig-11. *B* is the baseline of the geometrical parameters, D_0 is the scene ground distance, H_0 is the altitude of the flight. While the $Ø_0$ is the acquisition mean depression angle of the camera.

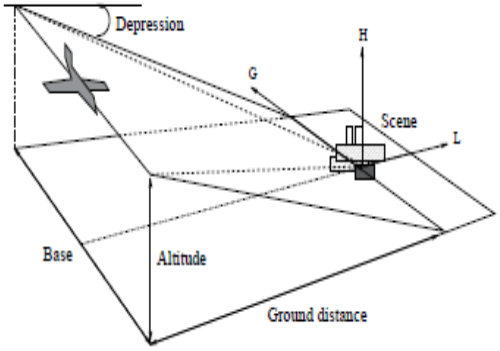


Fig.-11. Acquisition Geometrical Configuration

3. METHODOLOGY

In this study, we have adopted a Bag of Feature approach considering its simplicity in real time application. To estimate the altitude, the color space RGB is chosen, and then we de-correlate the luminance and chrominance.

With a given RGB image, it is converted to grayscale image using the RGB-to-grayscale conversion formula [1]. The mean values of the three components R, G and B generate a gray scale image:

$$\mathbf{Gray = (0.2126 \times Red^{2.2} + 0.7152 \times Green^{2.2} + 0.0722 \times Blue^{2.2})^{1/2.2}}$$ [1]

Furthermore, as we get the input image and convert it to grayscale we need to match the image we have in database as template to the input image. However, to process the matching we would chose an approach, The matching process moves the template image to all possible positions in a larger source image and computes a numerical index that indicates how well the template matches the image in that position. Moreover, to avoid the error due size change in close and far view we applied Spatial Pyramid approach as it start the matching with small template then increase the size for each cycle.

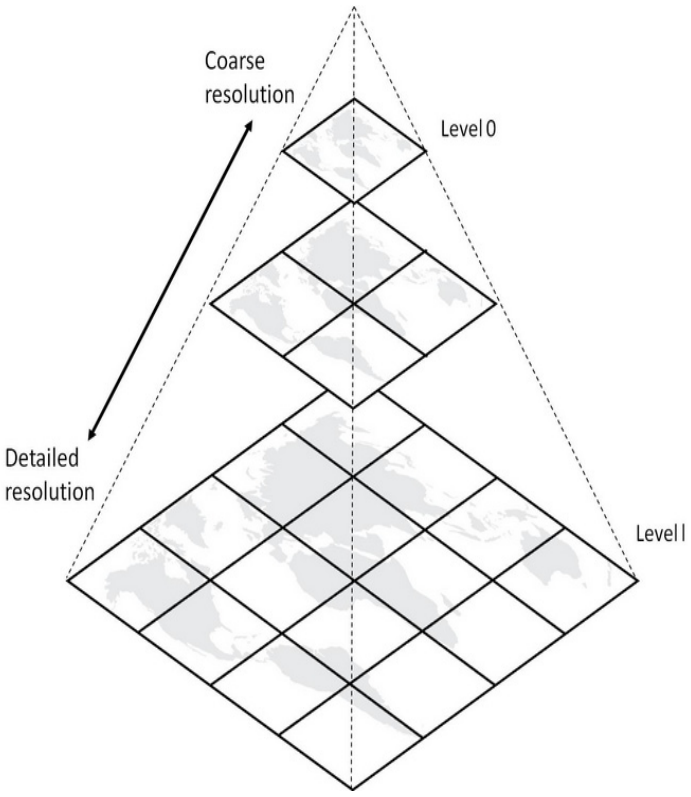


Fig-12. Spatial Pyramid starts with coarse resolution to avoid the huge processing of the detailed image.

Then feature extraction of this system uses SIFT & GIST which are the upgraded approaches of Histograms of Oriented Gradients Descriptors (HoG). However, this descriptor will be used to extract the features of the images then we do clustering to classify the images. However, there are many descriptors can be used or combined to work together in our system.

Fig-13. Image testing for HoG representation

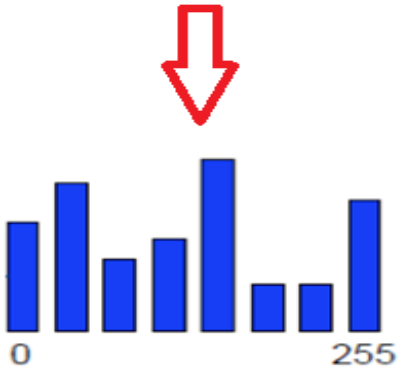


Fig-14. HoG representation for image.

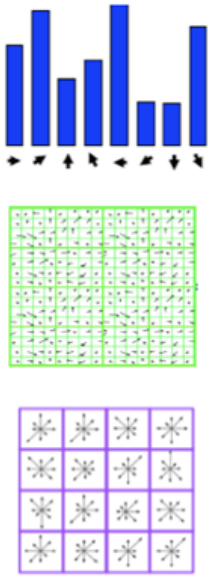


Fig-15. HoG representation is actually represented for group of pixel of image

Other descriptor apply the same operation of coveting the image pixels to vote to its color number as it described from (0 – 255), 0 for white ad 255 for black color. However, Divide the feature into log-polar bins instead of dividing the feature into square is the commune used approach. Furthermore, after extracting the features from the images we have to plot the features so we can classify the input, the system here used SVM to represent the fitting.

Fig-16. Fitting representation

Eventually, accuracy calculation has been done to show the system training and testing result.

$$\text{Accuracy} = \frac{\text{Number of correct classifications}}{\text{Number of instances in our database}}$$

4. EXPERIMENTS RESULTS

After the images have been trained by the system as tested with the testing images, the accuracy performed are 99.5% with 300 training images and 200 testing images

```
code vector: the 200 th image in class 1 is processing
code vector: the 199 th image in class 1 is processing
code vector: the 137 th image in class 1 is processing
code vector: the 198 th image in class 1 is processing
code vector: the 136 th image in class 1 is processing
code vector: the 197 th image in class 1 is processing
code vector: the 135 th image in class 1 is processing
code vector: the 196 th image in class 1 is processing
code vector: the 195 th image in class 1 is processing
testing phase is done!
Your accuracy rate is 99.5 %
```

Fig-17. Accuracy Test

5. CONCLUSION

With the 99.5% accuracy got from the training and testing of the data images, it can proof that our system is almost perfect for estimation. The main objective of this research project is to estimate the altitude from aerial images, many steps and research are taken to get it done perfectly with right estimation, but then to get the specific altitude is almost impossible, more time is needed to dig deeper into this research. So what is done in this research is that the training images are classified into certain range of altitude. For this research, the setting for class 1, the altitude range is from 0 - 1.5 meter, class 2 is from 2 - 4 meter, class 3 is from 5 - 7 meter, class 4 is from 8 - 10 meter, and lastly class 5 is range from 11 – 13meter height.

So from this setting, we manage to get almost perfect altitude estimation from aerial images with 99.5% accuracy. First we test for 50 training images, then we multiplied it to 100, then 150, the percentage of accuracy is increased with the increasing of training images, so at the end, the training images is increased up to 300 training images and the accuracy increased to 99.5% with 200 testing images.

As a conclusion, the main objective of this research that is to estimate the altitude from aerial images is successfully achieved with 99.5% accuracy.

REFERENCES

[1] A. Saxena, M. Sun, and A. Ng. 3-d scene structure from a single still image. In ICCV workshop on 3D Representation for Recognition (3dRR-07), 2007.

[2] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2005.

[3] J. Kim, V. Kolmogorov, and R. Zabih, "Visual correspondence using energy minimization and mutual information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nice, France, Oct. 2003, pp. 1033–1040.

[4] W. Liang, Novel Dense Stereo Algorithms For High-Quality Depth Estimation From Images, Theses and Dissertions-Computer Science. Paper 4, 2012

[5] "Inertial sensed ego-motion for 3d vision," J. Robot. Syst., vol. 21, no. 1, pp. 3–12, Jan. 2004.

[6] A. Saxena, S. Chung, and A. Ng. Learning depth from single monocular images. NIPS, 2005.

[7] A. Saxena, S. Chung, and A. Ng. 3-d depth reconstruction from a single still image. IJCV, Aug 2007.

[8] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: Transfer learning from unlabeled data. In Proceedings of the 24th International Conference on Machine Learning, 2007.

[9] H. Lee, A. Battle, R. Rajat, and A.Ng.Efficient sparse coding algorithms. NIPS 19, 2007.

[10] S. Geman, and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:721-741, 1984.

[11] J. Lobo and J. Dias, "Vision and inertial sensor cooperation using gravity as a vertical reference," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 12, pp. 1597–1608, Dec. 2003.

[12] S. Martial, B. Guy Le, and P.F. Sylvie. Height Estimation Using Aerial Side Looking Image Sequences, ISPRS Archives, Vol. XXXIV, Part 3/W8, Munich, 17-19. Sept. 2003

[13] C. Premebida, U. Nunes. Segmentation and Geometric Primitives Extraction from 2D Laser Range Data for Mobile Robot Applications, in *Proceedings of 5rd National Festival of Robotics Scientific Meeting (ROBOTICA)*, 2005: 17-25, Coimbra, Portugal.

[14] F. Lu, E. Milios, Robot Pose Estimation in Unknown Environments by Matching 2D Range Scans, *Journal of Intelligent and Robotic Systems*, Kluwer Academic Publishers, 18(3): 249- 275,1997.

[15] M. Sanfourche, G. Besnerais, and S. Foliguet. Height estimation using aerial side looking image sequences. ISPRS Archivs, Vol.XXXIV,Part3/W8, Munich, 2003.

[16] Q. Wei, "Converting 2d to 3d: A survey," Delft University of Technology, The Netherlands, Project Report, Dec 2005.

[17] M. Irani and P. Anandan, "A Unified Approach to Moving Object Detection in 2D and 3D Scenes," IEEE Transactions Pattern Analysis and Machine Intelligence, 20(6): p577-589, June 1998.