***Scene Understanding in Night-Time using SSAN Dataset***

**Abstract— Night vision greatly affects the efficiency of our vision which we**

**come across daily. Research work on night vision is very essential to solve the**

**social problems in the present scenario, but there is still a lack of database to**

**do research on night vision using deep-learning technique. Due to poor light**

**the object detection is a very tedious process. To overcome such hardships,**

**we collected the night vision datasets under various conditions. This work is**

**about scene understanding during night-time with IR-cameras. The feature**

**extraction from night videos is mainly affected by wavelength or the intensity**

**Of IR, illumination and distance factor. We proposed a novel algorithm exclusively for object detection during night time and we compare our algorithm with various yolo versions and we found that our night vision yolo performs better in detecting various objects like Male, Female, Car, bike, Van, Cycle during night time.**

**Keywords: Night vision, Night vision yolo, object detection, IR camera, SSAN datasets.**

1. **Introduction:**

Computer vision is used for the construction of meaningful description of physical objects from the obtained scenes. It is of two types day-time vision and night-time vision. In day-time the objects and its features can be easily extracted, but as a consequence of very low-light intensity it becomes difficult for the system to detect the objects and its features. In this work we mainly focus on the night vision system to improve the safety and security of people. Close circuit television (CCTV) is a system where the videos are transmitted in the closed form. With the help of CCTV we can record, monitor and control videos. Nowadays CCTV is mainly used for security systems in various sectors like banking, traffic, tourist places and educational sectors. However, in the videos obtained from IR-cameras, which are of low light intensity, research is more oriented towards enhancement of image, whereas the real challenge lies in object detection. Due to lack of benchmark datasets at night-time, research pertaining to this field has been superficial. There are many popular public object datasets such as PASCAL VOC[4,5], ImageNet[6,7], Microsoft COCO[8] and Exclusively Dark (ExDARK), which played a vital role in the object detection and recognition in low light environmentThe PASCAL VOC has a relatively big number of image datasets, comprising many variants that reflect genuine settings during a period of image data sets. Then ImageNet emerges in the year 2011,the consequence of which was the break-through of deep learning through the CNN, and a whole fresh wave of computer vision and machine learning activities was eventually triggered. While information sets proceed to expand in amounts, the type of information annotation creates a fresh task because natural annotators face difficulties in dealing with these figures. Then in 2014 comes the Microsoft COCO,which provided an extensive annotation, which covers a range of functions, including reconnaissance, segmentation and captioning, while not as big in figures as the ImageNet. While the advancement made with these datasets is remarkably evident, fewer than 2 percent of pictures recorded in low-light circumstances are important in these datasets. In addition, there are no datasets which offer the finest of our understanding specifically natural low-light pictures for object-focused functions.Exclusive Dark datasets are a set of 7,363 low-light pictures from very light to twilight settings with 12 categories of items (like PASCAL VOC) annotated both at picture category stage and on the local item bounding boxes. This dataset is a compilation of 10 distinct circumstances. In Indian scenario, the CCTV images are affected by high noise, poor illuminations, bad weather, so the above four datasets are not suitable for the Indian conditions during night time. First, we developed our own datasets called SSAN with IR camera which suits the Indian conditions for night time with various objects. Second, we developed our own neural network for night vision which satisfies various conditions. We have provided an object detection analysis for night time videos using night vision algorithms in both hand-crafted and learned features for understanding of night vision and its difference from vision with sufficient illumination. In chapter 2 we are going to discuss about the popular dataset and in chapter 3 is the proposed methodology and in chapter 4 is the results and discussion and in chapter 5 is the conclusion and references.

**SSAN datasets**

|  |  |  |
| --- | --- | --- |
|  | **A picture containing building, floor, outdoor, ground  Description automatically generated** |  |
|  |  |  |
|  |  |  |

**Figure 1:** Night-time CCTV IR footages with unequal lighting of SSAN datasets

Night vision system is an essential component of our everyday environment and has a significant impact on the efficacy of our sight. Night vision research has seen constant development, especially in image enhancement during dark time, but the database is still missing as a benchmark. Our dataset is for night vision objects, in which a dark night videos is classified as light if its illumination changes are either small or substantial. Night vision videos are taken with different condition with different illumination objects. The main challenge in night vision videos depend on the IR led light illuminations. In our CCTV system we are using the IR Led lights which can reach the max distance of 25 meters with the image sensor camera of 920pixel. We have considered various condition CCTV footage with various objects like Human, car, bike, bicycle, van & etc. For each object we have taken around 50 videos with frame rate of 25FPS. In figure 1: describes the night CCTV-IR images with various conditions like multiple objects, IR reflections, object shadows and IR illuminations. We have developed an exclusive benchmark dataset for night vision systems which will be useful for the night vision research scholars.

**Hand craft methods for night vision- drawback**

Hand craft feature extraction method is used for traditional machine learning approaches for object detection. Parameters which helps in image formation are reflectance, illumination,noise makes image clear and which makes object detection easier but in hand craft methods these 3 parameters makes it difficult to detect object which is the main drawback . So we developed an algorithm which overcomes these 3 parameters in detecting object and gives better result .

**Proposed Methodology**

There are 24 convolutional levels and two fully linked levels are present in the network architecture of this model. The convolutional strata extract the features while the fully linked strata estimate the location and probabilities of the boundary layers. At first, we divide the complete image into a panel grid of size n x n. Each grid cell associates with two bounding boxes and their respective category confidences, so we can identify a maximum of two items in a single grid cell. When an item occupies over one grid cell, we choose the middle cell to be the point of prediction for that item. A bounding box with no items has zero confidence value while a bounding box near an item has a confidence value congruous to the bounding box score.

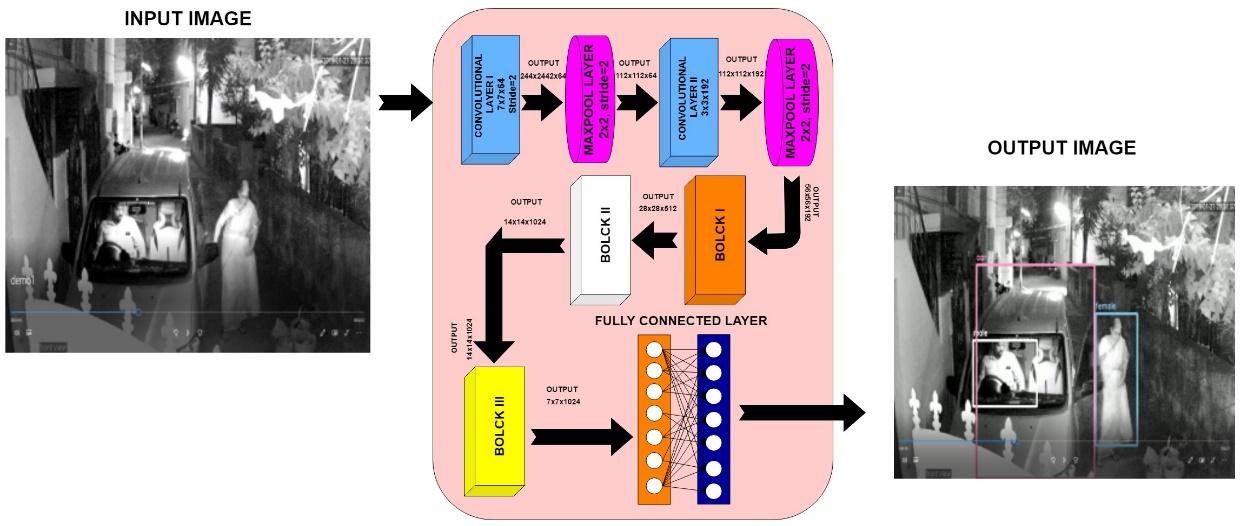


Figure 2: Proposed Night Vision YOLO Architecture

**Layer Description:**

Our model consists of 24 convolutional layers. Each layer is described in detail in the table below:

|  |  |  |
| --- | --- | --- |
| Name | Filters | Output Dimension |
| Convolution 1 | 7x7x64, stride=2 | 224x224x64 |
| Maxpooling 1 | 2x2, stride=2 | 112x112x64 |
| Convolution 2 | 3x3x192 | 112x112x192 |
| Maxpooling 2 | 2x2, stride=2 | 56x56x192 |
| Convolution 3 | 1x1x128 | 56x56x128 |
| Convolution 4 | 3x3x256 | 56x56x256 |
| Convolution 5 | 1x1x256 | 56x56x256 |
| Convolution 6 | 1x1x512 | 56x56x512 |
| Maxpooling 3 | 2x2, stride=2 | 28x28x512 |
| Convolution 7 | 1x1x256 | 28x28x256 |
| Convolution 8 | 3x3x512 | 28x28x512 |
| Convolution 9 | 1x1x256 | 28x28x256 |
| Convolution 10 | 3x3x512 | 28x28x512 |
| Convolution 11 | 1x1x256 | 28x28x256 |
| Convolution 12 | 3x3x512 | 28x28x512 |
| Convolution 13 | 1x1x256 | 28x28x256 |
| Convolution 14 | 3x3x512 | 28x28x512 |
| Convolution 15 | 1x1x512 | 28x28x512 |
| Convolution 16 | 3x3x1024 | 28x28x1024 |
| Maxpooling 4 | 2x2, stride=2 | 14x14x1024 |
| Convolution 17 | 1x1x512 | 14x14x512 |
| Convolution 18 | 3x3x1024 | 14x14x1024 |
| Convolution 19 | 1x1x512 | 14x14x512 |
| Convolution 20 | 3x3x1024 | 14x14x1024 |
| Convolution 21 | 3x3x1024 | 14x14x1024 |
| Convolution 22 | 3x3x1024, stride=2 | 7x7x1024 |
| Convolution 23 | 3x3x1024 | 7x7x1024 |
| Convolution 24 | 3x3x1024 | 7x7x1024 |
| Fully Connected I | - | 4096 |
| Fully Connected II | - | 7x7x30(1470) |

The first step towards YOLO explains how its output is encoded. The picture input is separated in the S x S cell grid. A grid cell is called "accountable" for anticipating each object present in the picture. This is the cell into which the object's core falls.

In each grid cell, bounding boxes B and C-probability are anticipated. The bounding box prediction has five components:(x, y, w, h, Confidence). The (x, y) coordinates are the center of the box, relative to the grid cell location (remember that if the center of the box does not fall inside the grid cell, this cell is not responsible for it). These coordinates are normalized between 0 and 1. The size of the (w, h) box is also normalized to [ –0, 1] relative to the size of the image.

The class probabilities Pr (Class(i) Object should also be predicted. This probability is determined by the grid cell of an object (see if the conditional probability means you do not know this). In practice, this means that the loss function does not penalize it for an erroneous prediction in the class if no object is present on the grid cell, as we shall see later. The network only predicts one class probability per cell, irrespective of box number B. This gives total probabilities of class S x S x C

Note that the design has been designed to be applied to the Pascal VOC dataset and authors used S=7, B=2 and C=20. The ultimate maps are therefore 7x7 and the input magnitude (7 x 7 x (2\* 5+ 20)) are also explained. This network may involve a tuning of layer sizes, with distinct grid sizes or different class numbers.

The sequences of 1x1 reduction layers and 3x3 convolutional layers were inspired by the GoogLeNet (Inception) model

The final layer uses a linear activation function. All other layers use a leaky RELU (Φ(x) = x, if x>0; 0.1x otherwise)

Loss Function:

(1)

where λ is a constant.

***𝟙 obj*** is defined as follows:

1, If an object is present in grid cell *ith* and the *j*th bounding box predictor is “responsible” for that prediction

0, otherwise

**Results and Discussion**

We ran this project on environment using CUDA CUDNN, and used a machine of the following specifications (intel i5 8th gen with 3.9Ghz,4 cores; NVIDIA GeForce GTX 1050 graphics card with 4GB of memory, 8 GB RAM). We trained this model with datasets containing 2485 images and had a testing dataset of 4 videos.

|  |  |  |
| --- | --- | --- |
| (a) | (b) | (c) |
| (d) | (e) | (f) |
| (g) | (h) | (i) |
| (j) | (k) | (l) |
| (m) | (n) | (o) |
| (p) | (q) | (r) |
| (s) | (t) | (u) |
| (v) | (w) | (x) |

Figure 3:{(a), (b), (c), (g), (h), (i), (m), (n), (o), (t), (u), (v)} - testing dataset {(d), (e), (f), (j), (k), (l), (p), (q), (r), (v), (w), (x)} - tested video using trained model

In our dataset we have 10 videos for each class, from which we split the video in rate of 10 frames/sec. So hence, 3000 images are generated. From that we took 2485 images for training dataset and remaining we kept it for testing for trained model.

The following way we trained the model:

1. First, pretrain the first 24 convolutional layers using the ImageNet 1000-class competition dataset, using an input size of 1080 X 720
2. Then, decrease the input resolution to 448x448
3. Train the full network for about 20 epochs using a batch size of 64, size
4. Learning rate schedule: for the first epochs, the average loss rate was slowly down from 78 to 6.8. Train for about 15 epochs and then start decreasing it.
5. Use random scaling and translation of information increases and adjust visibility and saturation uniformly.
6. We train the model upto 20 epochs to get the better result and the average loss function is 2.3.

It is important to modify the features of failure for stronger outcomes. Two items are remarkable:

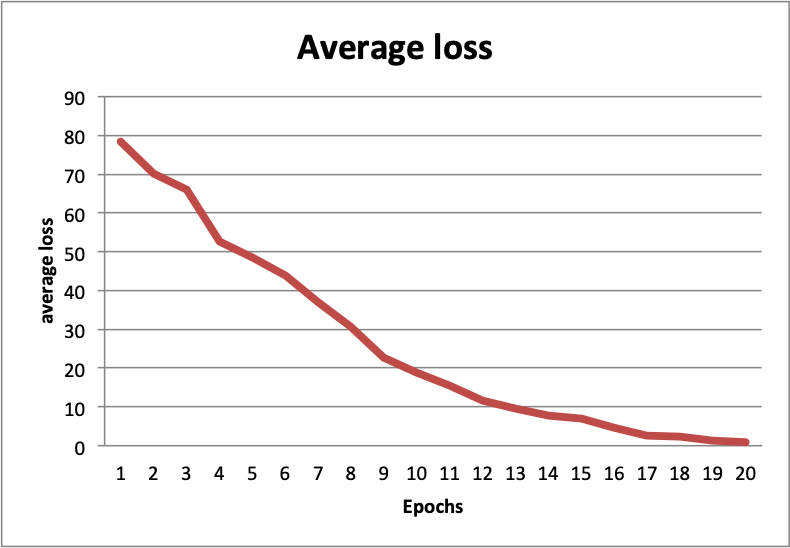
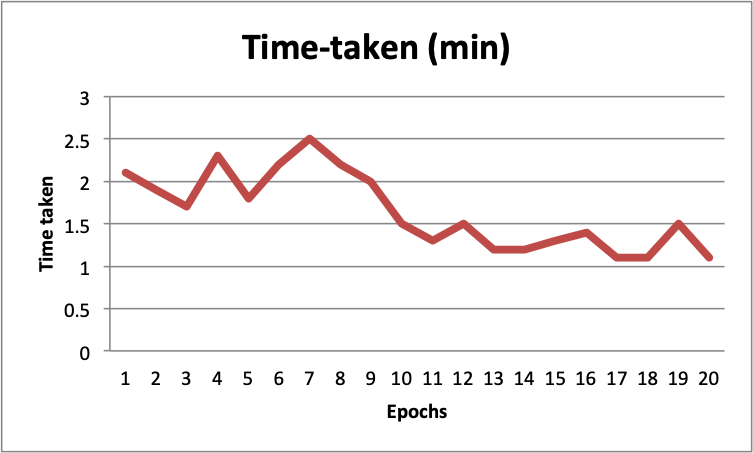
1. When multiple classes overlaps, labelling should be done properly.

2. Predict the bounding box size and height square root for penalizing errors separately in tiny objects and big objects.

Training datasets are used to train the model, which can detect various objects like male, female, car, van, bike, cycle. And we tested the model for testing videos, our model can detect objects based upon the wavelength of IR camera. And the advantage of this model is that it can detect objects lively, with less loss.

**Table 1: Comparison table with decreasing average loss with increasing epochs**

|  |  |  |
| --- | --- | --- |
| **Epochs** | **Average loss** | **Time-taken (min)** |
| 1 | 78.5 | 2.1 |
| 2 | 70.1 | 1.9 |
| 3 | 65.9 | 1.7 |
| 4 | 52.6 | 2.3 |
| 5 | 48.5 | 1.8 |
| 6 | 43.9 | 2.2 |
| 7 | 36.8 | 2.5 |
| 8 | 30.7 | 2.2 |
| 9 | 22.8 | 2 |
| 10 | 18.8 | 1.5 |
| 11 | 15.6 | 1.3 |
| 12 | 11.7 | 1.5 |
| 13 | 9.6 | 1.2 |
| 14 | 7.8 | 1.2 |
| 15 | 6.9 | 1.3 |
| 16 | 4.6 | 1.4 |
| 17 | 2.7 | 1.1 |
| 18 | 2.3 | 1.1 |
| 19 | 1.2 | 1.5 |
| 20 | 0.9 | 1.1 |

 ****

**Figure 4: Epochs vs Average loss**

**Figure 5: Epochs vs Time taken**

**Comparison of various yolo versions with our SSAN datasets**

|  |  |  |
| --- | --- | --- |
| **YOLO V1** | **YOLO V2** | **YOLO V3** |
|  |  |  |
|  |  |  |

Figure 6: **Comparison of night vision yolo with yolo versions**

Table 2: **Comparison of our algorithm with various yolo versions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SNO** | **algorithm** | **No. of frames** | **Threshold** | **Efficiency (%)** |
| 1 | yolo v1 | 606 | 0.7 | 72.5 |
| 2 | yolo v2 | 606 | 0.68 | 80.5 |
| 3 | yolo v3 | 606 | 0.75 | 86 |
| **4** | **proposed night vision yolo algorithm** | **606** | **0.72** | **92.25** |

**Conclusion**

This work introduces the SSAN data set in the hope of providing a complementary database for night time research work and encourages the community to address the challenges of long glossed night time environments, particularly in application-based research such as object detection.

Using this dataset, we analyzed in depth the computational behavior of the common, handmade and learned night time images in the context of the object detection and found some interesting information in them. The development of handcrafted characteristics was discovered to be primarily in bright circumstances, so that sound and absence of detailed information that often occur in nightlight pictures could not be properly addressed. Likewise, a modern denotational algorithm is not enough to manage the noise often occurring next to bright information.

In return, our study on learning characteristics using light and low-light information by coaching CNNs showed that, indeed, there should be enhanced numbers of low-light information to improve evening efficiency. We have also advanced to the understanding of the night time'' alters'' objects characteristics by visualizing the feature-vectors or exposure charts of a CNN, that the same item in shining and low-light results in very diverse characteristics.

Our research focuses on the assessment of features based on object detection and we think that in the night time domain more needs to be developed. Therefore, we hope that the SSAN database will be a useful database for future undertakings, both to further understand the vision behavior or to improve practical work performance at night. Our proposed algorithm gives the better result at 20 epochs with the average loss function is 2.3.

**References**

[1] Author, F.: Article title. Journal 2(5), 99–110 (2016).

[2] Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).

[3] Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).

[4] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. 111, 98–136.

[5] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88, 303–338.

[6] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015a. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. <http://dx.doi.org/10.1007/> s11263-015-0816-y.

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015b. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.