Chapter 3

Literature Review

After the preliminary study, we can focus on presenting the state of the art, describing the accessible solutions and how they fit into the project's software. In this chapter, we will address the field of multi-sensor image fusion in literature and present the proposed methods with its results. Later, we will examine how object detection and classification tasks were approached in literature and explain in detail various deep learning concepts and architectures and how they have been used.

3.1 Multi-Sensor Image Fusion

In this section, we will explain what is multi-sensor image fusion, when is it required and how it can help in with the project. Then, we will review in detail the literature of image fusion techniques and the solutions that are related to the current task and how they can be utilized to solve it.

3.1.1 Definition

The image fusion process is characterized as gathering all the significant information from different images, and combining them into fewer images, generally a single one. This single image is more enlightening and accurate than any single source image, and comprises all the essential information. The reason of image fusion isn't only to decrease the sum of information but moreover to develop images that are more relevant, instructive and comprehensible for the human and machine vision.

3.1.2 Visible and Thermal Image Fusion

Trusted vision in challenging light conditions is one of the critical requirements of phenotyping. Within the ultimate decade, thermal cameras have finished up more easily available

to a bigger number of analysts. This has brought about in different considers which confirmed the benefits of the thermal cameras in limited visibility conditions.

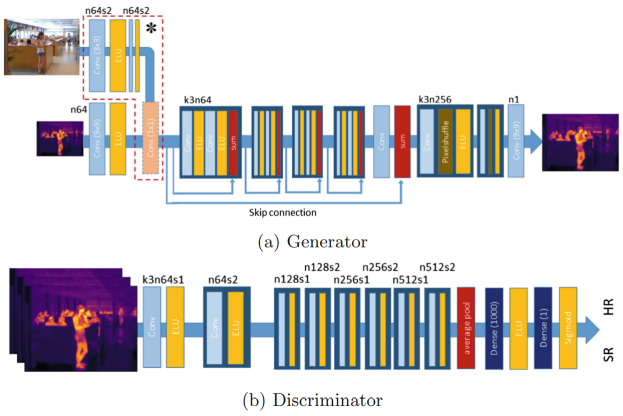
In digital images, what we perceive as details greatly depends on the image resolution. The higher the resolution the more accurate the estimation. The color camera performs well in great brightening conditions, giving a wealth of colors and detailed data however poor light conditions compromise its effectiveness.

The thermal cameras work well in constrained visibility conditions. They are utilized in low-resolution and in low contrast. However it is ineffective under direct bright sunlight. Therefore visible data is found to be insufficient and thermal images have become a common tool to overcome these problems. Increase the overall system robustness to varying lighting conditions since it ensures that information from at least one sensor is solid under all lighting conditions.

In this subsection, we will list the solutions proposed in the literature in order to fuse RGB images with infrared images.

- **Multimodal Sensor Fusion in Single Thermal Image Super-Resolution[Feras Almasri & Olivier Debeir, 2019]:**

  This research paper aims to enhance the thermal image resolution via the integration of high-frequency information from the visual image. The authors of the paper got inspired by [Ledig et al,2016] that present Super-Resolution Resolution Generative Adversarial Network (SRGAN) architecture. They were also influenced by different network architecture schemes, optimization functions, learning procedures and up-sampling methods proposed in a group of papers. They propose a new deep residual network by the contribution of Generative Adversarial Network (GAN)Structural Similarity index based model as illustrated in the figure n°3.1.



(a) Generator

(b) Discriminator

Figure 3.1: The Proposed Solution Architecture

They compare their model with the state of the art models using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure(SSIM) metrics and find out interesting outcomes. However, they discover that the main artifact is the device design or the

displacement of the device or the object.

- **Deep Visible and Thermal Image Fusion for Enhanced Pedestrian Visibility[Ivana Shopovcka et al, 2019]**

  The authors propose a learning-based approaches for thermal and visible image fusion that concentrates on producing fused images with high visual similarity to RGB images in pedestrian regions. This research paper aims to create intuitive and natural images that may be more instructive than a RGB images in challenging visibility conditions. The proposed method is a Convolution neural network architecture as shown in the figure n°3.2.

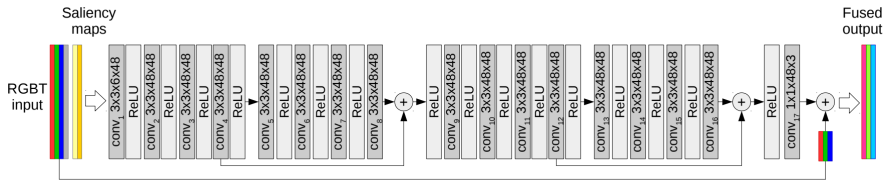


Figure 3.2: The Architecture of the Proposed Fusion Network

The authors evaluate the generated images by assessing object detection on it using objective metrics. They test object detection task using these two models: YOLO v2 and Faster R-CNN. YOLO v2 gives the top outcomes on the proposed images as in shown in the figure n°3.3.

Fusion Method	MI	$Q^{AB/F}$	IE	AG	SF	Miss Rate (%) Day/Night	
						Faster R-CNN	YOLO v2
No fusion (RGB only)	n/a	n/a	n/a	n/a	n/a	66.28/68.58	57.26/55.95
CNNs IR + VIS [5]	2.00	0.72	**7.47**	6.39	11.16	74.30/74.25	62.58/59.52
Hybrid MSD [6]	1.90	0.69	7.35	**7.10**	**12.3**	74.93/75.59	61.37/60.85
Image Fusion with Guided Filt. [7]	2.32	0.60	7.10	4.36	8.95	85.85/77.71	73.27/59.23
ResNet50 + CA [5]	1.73	0.61	6.81	3.80	7.50	75.36/75.50	64.56/60.11
Image Fusion using Deep Learning [9]	1.93	0.59	6.76	3.73	7.41	75.65/73.74	63.48/60.73
Structure-Aware Image Fusion [10]	3.48	**0.73**	7.45	6.49	11.60	75.75/71.08	64.44/57.82
Image Fusion using CBF Codes [11]	1.92	0.68	7.25	5.85	10.56	78.31/77.23	68.57/63.27
Image Fusion based on FPDE [12]	1.97	0.63	6.51	4.51	8.02	74.05/76.74	59.68/58.19
Two-Scale Image Fusion w/Sal. [13]	1.53	0.68	7.00	6.14	10.73	73.64/70.57	64.03/58.66
MSVD Image Fusion [14]	1.79	0.60	6.77	4.93	8.97	80.01/75.38	65.78/50.49
Proposed	**4.16**	0.70	7.23	5.67	11.19	**52.07/43.25**	**55.85/44.12**

Figure 3.3: Architecture of the Proposed Fusion Network

- **RGB Guided Thermal Super-Resolution Enhancement[Feras Almasri & Olivier Debeir, 2018]:**

  The authors worked in improving object detection algorithms in different conditions of illumination and occlusion. In this paper, the authors propose a deep learning model aiming to enhance the thermal image resolution guided by RGB images using GAN based model. They introduce the architecture in the figure n°3.4.

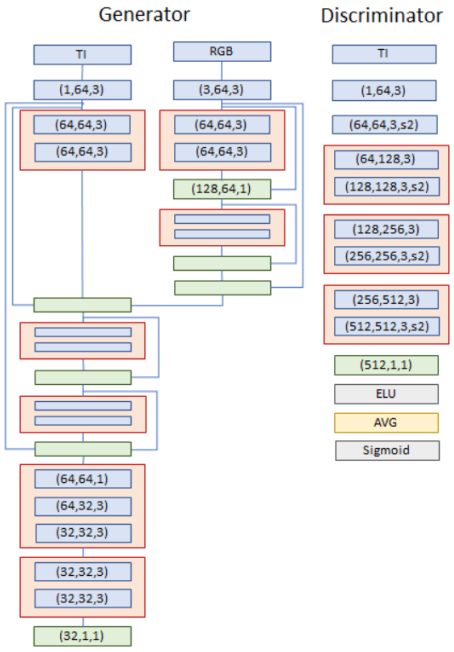


Figure 3.4: Architecture of the Proposed Fusion Network

The writers assess the performance of their model by comparing it with other models as in the figure n°3.5.

	bi-cubic	SRCNN-RGB	TIGAN-RGB	SRCNN	TIGAN
PSNR	45.471	51.864	50.578	**52.366**	51.826
SSIM	0.832	0.939	0.932	0.943	**0.944**

Figure 3.5: Quantitative Evaluation on Different Models

The results show that SRCNN gives satisfying outcomes but misses high-frequency details that arise in blurred pictures adding to over smoothed textures. However, GAN-RGB model learns more effectively the data distribution and produces sharper textures. The authors finds out that the foremost artifact consists in the displacement between the device and the objects.

- **IFCNN: A general Image Fusion Framework based on Convolutional Neural Network[Yu Zhanga et al, 2020]:**

  This paper proposes a general image fusion framework on the basis of the convolutional neural network, known as IFCNN. The authors were inspired by the transform-domain image fusion algorithms and used two convolutional layers to extract the salient image features from numerous input pictures. Later, the convolutional features of numerous pictures are fused by an relevant fusion rule; elementwise-mean, elementwise-max and

elementwise-min. Eventually, the fused features are reconstructed by two convolutional layers to create the instructive fusion image. Figure n°3.6 illustrates the proposed architecture.

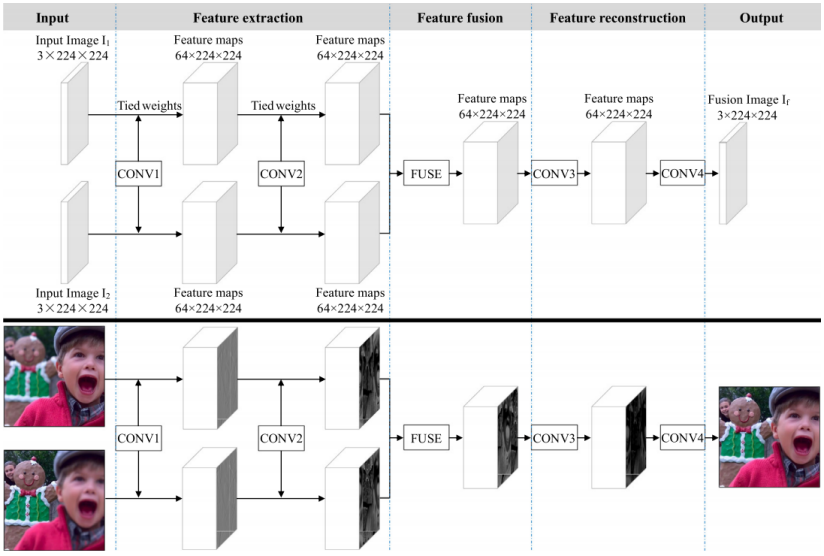


Figure 3.6: General Image Fusion Framework Based on Convolutional Neural Network : IFCNN

The experimental outcomes demonstrate that the proposed model illustrates higher generalization capability than the available image fusion models for fusing diverse types of images, including infrared-visual, multi-focus, multi-exposure and multi-modal medical images. Besides, the outcomes check that the proposed model has accomplished equivalent or even better outcomes compared to the state-of-the-art image fusion models on four types of image datasets.

- **DenseFuse: A Fusion Approach to Infrared and Visible Images [Hui Li & Xiao-Jun Wu, 2019]**

  The authors introduced a new deep learning architecture for visible and infrared images fusion task. Contrary to typical convolutional networks, the proposed encoding network is joined by convolutional layers, dense block in which the output of each layer is linked to every other layer and fusion layer as shown in the figure n°3.7. They utilized this architecture to get more valuable features from input images in encoding stage. And two fusion layer are planned to fuse these features. Eventually, the fused image is reconstructed by decoder.

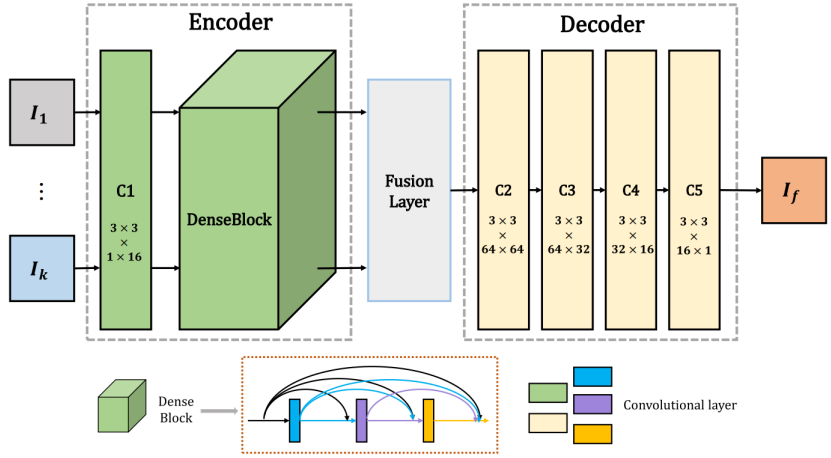


Figure 3.7: Image Fusion Framework Based on Convolutional Neural Network : IFCNN

The authors compared the proposed fusion method with existing fusion methods using objective and subjective assessment as shown in the figure n°3.8.

Methods			En	Qabf [25]	SCD [26]	FMI_w [27]	FMI_{dct} [27]	$SSIM_a$	MS_SSIM [28]
CBF [22]			6.81494	0.44119	1.38963	0.32012	0.26619	0.60304	0.70879
JSR [9]			6.78576	0.32572	1.59136	0.18506	0.14184	0.53906	0.75523
GTF [23]			6.63597	0.40992	1.00488	0.41004	0.39384	0.70369	0.80844
JSRSD [24]			6.78441	0.32553	1.59124	0.18502	0.14201	0.53963	0.75517
CNN [13]			6.80593	0.29451	1.48060	**0.53954**	0.35746	0.71109	0.80772
DeepFuse [15]			*6.68170*	0.43989	*1.84525*	0.42438	*0.41357*	*0.72949*	**0.93353**
ours	Addition	$\lambda = 1e0$	6.66280	0.44114	**1.84929**	*0.42713*	**0.41557**	**0.73159**	*0.93039*
		$\lambda = 1e1$	6.65139	0.44039	**1.84549**	*0.42707*	**0.41552**	**0.73246**	*0.92896*
		$\lambda = 1e2$	6.65426	*0.44190*	**1.84854**	*0.42731*	**0.41587**	**0.73186**	*0.92995*
		$\lambda = 1e3$	6.64377	0.43831	1.84172	*0.42699*	**0.41558**	**0.73259**	*0.92794*
	l_1-norm	$\lambda = 1e0$	**6.83278**	**0.47560**	1.71182	*0.43191*	0.38062	0.71880	0.85707
		$\lambda = 1e1$	6.81348	**0.47680**	1.71264	*0.43224*	0.38048	0.72052	0.85803
		$\lambda = 1e2$	**6.83091**	**0.47684**	1.71705	*0.43129*	0.38109	0.71901	0.85975
		$\lambda = 1e3$	**6.84189**	**0.47595**	1.72000	*0.43147*	0.38404	0.72106	0.86340

Figure 3.8: The Average Values of Quality Metrics for 20 Fused Images

3.1.3 2D and 3D Image Fusion

2D images have progressed to be reasonably accurate beneath controlled conditions. But it have proved that their performance decreases significantly when pose or brightening variations are present in pictures. Moreover, the characteristics of 3D point clouds and 2D digital images are thought to be complementary. In order to improve the image quality under these conditions, merging 2D image and 3D image should be studied.

In this subsection, we will list the solutions proposed in literature in to fuse 2D images with 3D images.

- **Mobile thermal mapping for matching of infrared images with 3D building models and 3D point clouds[L. Hoegner & U. Stilla, 2018]:**

  The authors propose to match a TIR-image based 3D point cloud and a high resoluted RGB-image based 3D point cloud in order to appoint thermal intensities to the dense RGB point cloud. The scheme of the proposed method is presented as shown in the figure n°3.9.

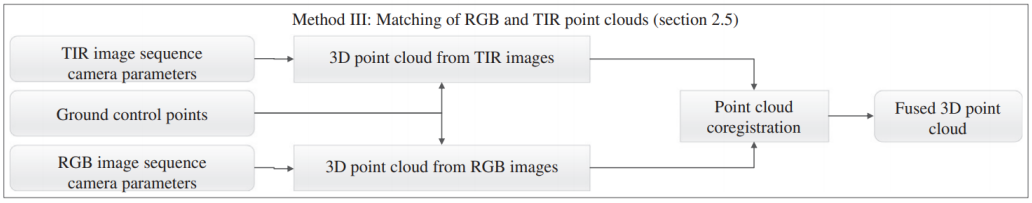


Figure 3.9: Scheme of the Presented Method

The 3D point cloud from TIR images and the 3D point cloud from RGB images are matched thanks to the pre-orientation of the two-point clouds. The fixed orientation is exchanged to the orientation parameters of the TIR-images. The ultimate image orientations are utilized to extract thermal textures in order to achieve the 3D building model.

The results showed that the main limitating factor for the quality and the accuracy of the point cloud projection was the reliable and accurate camera orientation.

- **Edge Extraction by Merging 3D Point Cloud and 2D Image Data [Ying Wang et al, 2013]:**

  This paper introduces a novel method to extract 3D edges. The idea consists in merging the edge data from a point cloud of an object with its matching digital pictures. The proposed method is attempted to take advantage of image processing and both edge processing and analysis of point clouds to constitute the edge characteristics in 3D with augmented accuracy. After applying data pre-processing on raw images, an edge extraction is employed upon the Canny edge detection algorithm. Later, a pixel data mapping process is executed to correspond 3D point cloud pixels with 2D image pixels. Next, 2D edge data are combined in a 3D point cloud. With the application of the Point Cloud Library (PCL), edge points in the obtained image are converted to 3D point type in order to determine the edges in the 3D point cloud. The workflow is presented as illustrated in the figure n°3.10.

  The results showed that thanks to 2D-3D pixel mapping process, 2D edge pixels are properly fused into the point cloud. At the same time, the 3D edges do not correspond to the edge features of the object due to the difference of 2D and 3D processing in edge detecting accuracies.

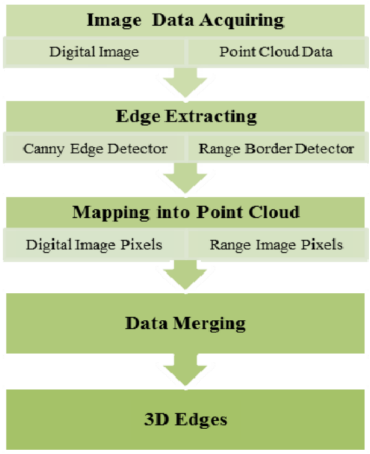


Figure 3.10: Proposed Workflow

- **Infrared and 3D Skeleton Feature Fusion for RGB-D Action Recognition[Alban main de Boisserie & Rita Noumeir, 2020]:**

  This paper propose a new network to merge skeleton and infrared data as shown in the figure n°3.11. In order to extract features from skeleton data and visual features from videos, a pre-trained 2D convolutional neural network and pre-trained 3D CNN are respectively used. Later, both feature vectors are merged via a multilayer perceptron (MLP). The researchers tested various methods for fusion with the aim to find out the best method.

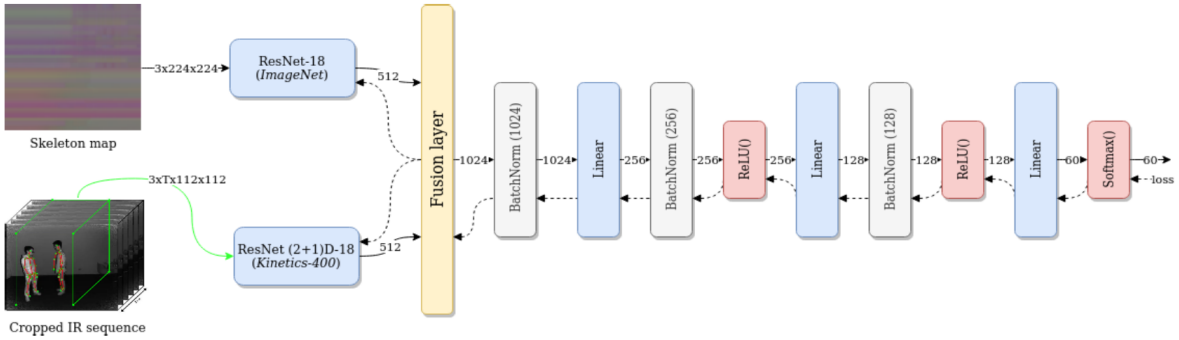


Figure 3.11: Proposed Model Architecture

The researchers compared the accuracy of their model with the state-of-the-art models and found that they achieve better results as illustrated in the figure n°3.12.

Method	Pose	RGB	Depth	IR	CS	CV
Lie Group [53]	X	-	-	-	50.1	82.8
HBRNN [9]	X	-	-	-	59.1	64.0
Deep LSTM [38]	X	-	-	-	60.7	67.3
PA-LSTM [38]	X	-	-	-	62.9	70.3
ST-LSTM [32]	X	-	-	-	69.2	77.7
STA-LSTM [47]	X	-	-	-	73.4	81.2
VA-LSTM [64]	X	-	-	-	79.2	87.7
TCN [23]	X	-	-	-	74.3	83.1
C+CNN+MTLN [21]	X	-	-	-	79.6	84.8
Synth. CNN [33]	X	-	-	-	80.0	87.2
3scale ResNet [27]	X	-	-	-	85.0	92.3
DSSCA-SSLM [39]	-	X	X	-	74.9	-
[36]	X	-	X	-	75.2	83.1
CMSN [67]	X	X	-	-	80.8	-
STA-HANDS [2]	X	X	-	-	84.8	90.6
Coop CNN [58]	-	X	X	-	86.4	89.0
ST-GCN [62]	X	-	-	-	81.5	88.3
DGNN [41]	X	-	-	-	89.9	**96.1**
Pose module - PA	X	-	-	-	81.9±0.28	89.6±0.53
IR module - CPA	-	-	-	X	90.4±0.79	93.8±0.46
FUSION - CPA (sum)	X	-	-	X	**91.8±0.40**	**94.9±0.39**

Figure 3.12: Comparison of the Proposed Model to the State of the Art

3.2 Object Recognition and Classification

In this subsection, we will give a brief overview of the traditional object identification algorithms that use Convolutional Neural Networks (CNNs), as well as their latest developments. Furthermore, techniques for merging CNNs using region proposal generators should be explored extensively.

3.2.1 Object Recognition

The fastest-rising viewpoints of computer vision is object recognition. Owing to progessed computing performance and huge publicly accesible datasets, a machine can now smoothly exceed a normal person in object detection. In spite of the the popularity of the earlier cited computer vision applications has risen sharply within the recent ten years, they have a prolonged history beginning within the early 1960s.

3.2.1.1 Classical Methods

Classic object recognition techniques were mostly centered on the recovery of feature descriptors. Tantamount to Histograms of Orientated Gradients (HOG)[Dalal, Navneet and Bill Triggs, 2015] or scale-invariant feature transform (SIFT) [Lowe David G, 2015], it was verified by a classification algorithm, as SVM, pursed by the application of deep neural networks for object identification. These traditional methods are concisely laid out in this section.

- **Histograms of Oriented Gradients for Human Detection**

  Histogram of Oriented Gradients (HOG) is a deep feature descriptor for pictures. The basic idea is that the local intensity gradients could be utilized to characterize the geometry of the frame. The HOG method extricates data about gradients to catch these forms in a zone (“Histogram of oriented gradients”). HOG divides a picture into a set of interlinked

spatial regions generally. Every region is subsequently divided into n $\times$ n cells; each cell holding a foreordained amount of gradient orientation bins over n pixels.

Each pixel within the cell gives a rate for a gradient orientation bin which is corresponding to the magnitude of the gradient at that pixel. After computing the intensity gradient for each pixel within the cell, the direction of the gradient in various cells is measured. HOG employs a 'sliding window' technique to extricate features and categorize zones from sub-images at various sizes and perspective proportions. The united feature vector of a linear classification for example SVMs heads towards the extreme yield when these slider windows are wound into a thickly covering matrix with HOG descriptors. HOG needed stability with regards occlusion and distortions. The HOG procedure's accuracy and detection speed were without doubt costly for numerous applications.

- **Object Recognition from Local Scale-Invariant Features** The concept of Scale Invariant Feature Transform (SIFT) which may be a picture descriptor boardly used for the individual/face detection and classification. Owing to its invariancy of translating, orientations and scaling transformations, The SIFT classifier appeared exceedingly efficient for object recognition withinn the real environment. A SIFT descriptor's task may well be partitioned into the following phases.

  1. The first stage is to recover pertinent or areas of interest originated from labeled greyscale images. These essential points are situated in 2D space where the vacillation of the signal outperforms a particular limit. SIFT makes an organizing element to extricate these essential points. This scaling structure is given by sewing together a grouping of Gaussian-blurred frames. After getting the essential points, the histogram of gradient directions in a constrained zone around each essential point is calculated by a operational vector.
  2. The second phase consists in the development of a feature vector-like descriptor using the recognizing characteristic of SIFT.
  3. Within the third phase, poor essential points like low contrast areas and edges are expelled and an assembly of orientation histograms for said persisting essential points is computed. Finally, based on the scaling and rotation in-variance built up, a change is utilized to locate groups from a single item. The probability of a particular feature vector matching an item within the frame is later computed. The assessment is performed as a least-squares reliable execution to the affine change factors.

- **OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks**

  Sliding window processing approach is essentially utilized in early tries at object identification thanks to CNN. These concepts have proven to be very viable in a diversity of fields, eminently vehicle, face adding to text detection. Thanks to the development of digital region-based object detectors, the authors proposes an innovative model by

combining image classification and regression operations giving as a result a Single Convolutional Neural Network.

The essential purpose was to create categorization at different regions on a diversity rate of a frame, in a slicing window way and to induce bounding-box regression to encompass an object more firmly.

This approach starts by training a CNN model for image classification. Later, the classification algorithm shapes are substituted by a label-wise regression structure that is trained to enhance the edges of the boxes at each picture plane. Ultimately, the predicted bounding boxes and resulting class rankings are aggregated , using a greedy merging method.

In spite of the early victory, such system have some downsides. The main drawback is the reiteration of the refined slide , which essentially raised the changing time. Nevertheless, the approach of object localization in this work has been enhanced within the current object detection job.

3.2.1.2 Region-Based Convolution Neural Network (R-CNN)

- **Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation :**

Region-based Convolution Neural Network (R-CNN) model includes three components.

The primary component generates region thoughts by using an external algorithm to search the input picture looking for conceivable objects. There are multiple ways of creating region proposals, counting objectness, ,Selective Search, Obliged Parametric Min-Cuts (CPMC),Category Independent Object Proposal, Multi-Scale Combinatorial Gathering and others.

The second module of R-CNN is the Selective Search algorithm; the foremost frequently utilized method for creating region suggestions. Then, these regional propositions are divided together to fit the input of an enormous, multi-layered CNN. This enormous CNN recovers one picture from each region with a fixed-length specific vector.

The third module comprises of a set e of linear support vector machines (SVMs) used in category-specific applications. The functional vectors of this method are feed into the classifier with the use of the Non-Maximum Suppression (NMS). A linear regressor is provided with the feature vector with its expected border boxes of the objects.

In figure 3.1 we show three modules in R-CNN.

Although R-CNN was instinctual, it had some downsides. One of the main downsides was the fixed input dimension for regional proposals, which were exceeding the predefined dimension before the CNN processing by disturbing picture areas. The second problem of R-CNN is its high computing expense.

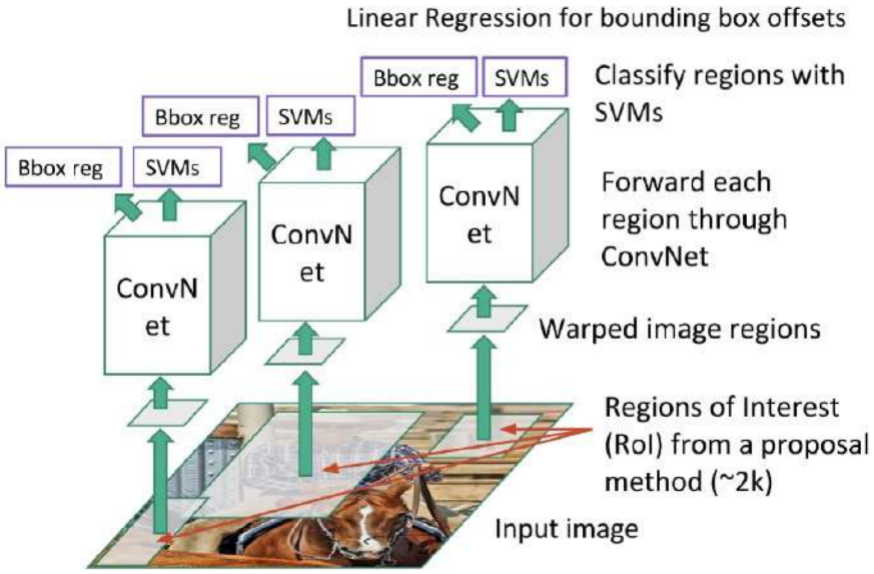


Figure 3.13: Region-Based Convolution Neural Network

3.2.2 Classification

In this section, we will explore and explain the main concepts that deal with the problems of plant classification or something similar. We will analyze the usefulness of these methods and whether or not they can be utilized. Classification methods in computer vision problems can be split into major branches. The first branch utilizes image processing techniques in order to extract numeric features that encode the object we are trying to classify and then these features are fed to Machine Learning classifiers. The second branch utilizes Convolutional Neural Networks (CNN) for encoding and classifying the images. In this section, we will focus on the first branch and we will explore in detail the second branch in the next subsection.

3.2.2.1 Using Image Processing Techniques

Feature extraction focuses usually on 3 main axes: morphological, color and texture. Morphology features encode the form and structure of the seeds like the size, shape and axis lengths. Color features represent the colors of the image in the case of both RGB images and Gray Scale images. They are usually extracted from the histograms of the pixel values. Texture features correspond to textural characteristics of the seeds. The Gray level co-occurrence matrix and Gray level run length matrix are two examples of these features. Feature extraction is a long and iterative process that requires a good intuition and domain knowledge from the researcher. Choosing the right features depends heavily from the problem as well as the classifier.

- **Digital image processing techniques for detecting, quantifying and classifying plant diseases [Jayme Barbedo, 2013]:**

Treat plant diseases classifications utilizing image processing and later examined them. In spite of the fact that most of the approaches and mentioned papers concentrate on distinctive tasks, the techniques applied and their outcomes provide significant understanding into which algorithms to test first together with the top approaches used in 2013 in this specific field. The writer used a group of techniques for each task type:

- **Detection:**
  - Dual-segmented regression analysis
  - Thresholding
  - Neural networks
- **Quantification:**
  - Thresholding
  - Color analysis
  - Fuzzy logic
  - Knowledge-based system
  - Region growing
  - Third party image processing packages
- **Classification:**
  - Neural networks
  - Support vector machines
  - Fuzzy classifier
  - Feature-based rules
  - Color analysis
  - Self organizing maps
  - Discriminant analysis
  - Membership function

- **Application of Image Processing in Agriculture: A Survey[Anup Vibhute & S.K.Bodhe, 2012]**

  This paper could be an overview on the common utilize of image processing within the agricultural field. The writer records different algorithms for classification like Artifical Neural Systems, KNN and PCA. The author gives various features that may be extracted and utilized. However the paper does not give a definitive conclusion but it have showed that image processing may be valuable for facilitating and speeding up the farmers tasks except that none of the cited approaches were justified in detail.

3.2.2.2 Using Deep Learning Techniques

- **Inference of Plant Diseases from Leaf Images through Deep Learning [harada Mohanty et al, 2016]:**

  The writers of this paper point to evaluate the performance of Deep Learning methods within leaf image classification task. For this purpose, they utilized three forms of a leaf image dataset ; the first involves segmented images, the second grayscale images and the last color images. Utilizing this approaches, adding to diverse train-test divides of the dataset, they are capable to evaluate whether these algorithms are able of learning to differentiate leaf diseases.

  They evaluate this hypothesis via 2 distinctive CNN architectures. AlexNet[Alex Krizhevsky et al, 2012] is the first architecture which revolutionized the utilize of CNNs in classification tasks by winning on September 30 2012,the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). GoogleNet [Christian Szegedy et al, 2015] is the second architecture used, also called as InceptionV1. They trained the two architectures from scratch and utilized transfer learning resulting in 60 configurations.

  The ultimate outcomes demonstrated that GoogleNet with transfer learning is better than AlexNet. The writers examine also the utilize of classical feature extraction approaches and they share the vision that hand-engineered features presented many bias and need a huge amount of time to be well specified for a particular task.

- **Deep machine learning provides state-of-the-art performance in image based plant phenotyping [Michael Pound et al, 2016]:**

  This paper utilizes 2 distinctive CNNs to fix distinct tasks; the first consists in a simple classification task whereas the second consists in a localization task. They trained, in the first task, an 11-layers-sized CNN architecture (2CL-1PL-2CL-1PL-2CL-3FCL) that they named Root CNN. Fot the second task, they utilized a CNN architecture within 15 layers (2CL, 1PL, 3CL, 1PL, 3CL, 1PL, 1CL, 3FCL) that they named Shoot CNN where PL, CL and FCL signify respectively 2x2 Pooling Layer, 3x3 Convolution Layer and Fully Connected Layer. The authors utilized the results of these models to perform a QTL analysis which are regions of the DNA that connect with phenotypic changes.

Conclusion

In this chapter, we detailed the state of the art, we explained the multi-sensor image fusion concepts and we reviewed object detection and classification papers. From these image fusion articles, we can conclude that the best approaches are proposed [Ledig et al,2016] and [Yu Zhanga et al, 2020]. As for the classification task, it is obvious that YOLO approach fits our requirements.