A Pose based Object Recognition Model for Improving Learning Time and Accuracy

Ankur Chauhan
Jaipur National University
Jaipur

Sanjay Kumar, Ph.D.
Jaipur National University
Jaipur

ABSTRACT

Now in these days the computational domain contributes in a different intelligence applications such as decision making, data analysis, and face recognition and pattern detection. These applications are supporting in various real world applications. In this paper, the pattern analysis and pattern discovery task is discussed for object recognition application. Object recognition is a computational process where using the visual features are utilized for approximating the actual real world objects. In literature there are a number of object recognition models are available, those are promises to provide accurate object detection. But most of them are only produces 40-50% accurate results. In this paper basically different object recognition models are discussed which are providing guidelines for obtaining accurate model. In addition of that this paper addresses the real world issues which are required to involve for future object recognition model.

Keywords

Object recognition, review, accurate modeling, issue and challenges, proposed model.

1. INTRODUCTION

Machine learning and artificial intelligence is a subject which contributes on developing the knowledgeable system. The machine learning based systems are capable to make decisions, helps on problem analysis, searching for domain specific solutions. Among them object recognition is a new and interesting domain of research and development. In the object recognition the objects and their visual patterns are learned by machine learning system, and based on previous knowledge the objects are recognized by these algorithms.

Therefore, we can say the object recognition models are prepared in two major modules first training with object examples and then recognition of these real world objects. In real world the object can be classified as movable and static objects. Movable objects are changes their positions with the time slices and the static objects are placed in a specific place and not changing their place automatically without any external forces. Thus the ***mobility*** of object is a significant attribute in object recognition.

Each object having their dimensional information, mean to say each object having a ***volume*** means height, width and length. Additionally an object is also recognized by their colure and texture. Therefore the objects can be recognized using their physical or visual properties.

In this paper we are focused on the finding the object properties and their estimation techniques. In addition of that how these features are calculated a study is performed finally a newer system is introduced which is further implemented and performance of the system is evaluated with real world objects.

2. LITERATURE SURVEY

In this section recent contributions and newer approaches are discussed which are providing guidelines for enhancing the object recognition process.

R. Lefort et al [1] addresses the inference of probabilistic classification models using weak supervised learning. The main contribution of this work is the development of learning methods for training datasets consisting of groups of objects with known relative class priors. Training information is given as the presence or absence of object classes. Generative and discriminative classification methods are conceived and compared for weakly supervised learning, as well as a nonlinear version of the probabilistic discriminative models is also provided. Additionally, considered models are evaluated on standard datasets and an application to fisheries acoustics is reported. The proposed proportion based training is demonstrated to outperform model learning based on presence/absence information and the potential of the non-linear discriminative model.

Zhangzhang Si et al [2] presents a framework for unsupervised learning of a hierarchical reconfigurable image template—the AND-OR Template (AOT) for visual objects. The AOT includes: 1) hierarchical composition as “AND” nodes, 2) deformation and articulation of parts as geometric “OR” nodes, and 3) multiple ways of composition as structural “OR” nodes. The terminal nodes are hybrid image templates (HIT) [3] that are entirely creative to the pixels. Author shows that both the structures and parameters of the AOT model can be learned in an unsupervised way from images using an information projection principle. The learning algorithm includes of two steps: 1) a recursive block pursuit method to learn the hierarchical dictionary of primitives, parts, and objects, and 2) a graph compression method to reduce model structure for better generalizability. They examine the factors that influence how well the learning algorithm can recognize the underlying AOT. And propose a number of ways to calculate the performance of the learned AOTs through both synthesized instances and real-world images. Given model advances the state of the art for object detection by improving the accuracy of template matching.

Juergen Gall et al [4] introduces Hough forests which are random forests personalized to perform a generalized Hough transform in a proficient way. Compared to earlier Hough-based systems like implicit shape models, Hough forests advance the performance of the generalized Hough transform for object detection on a categorical level. At the same time, their flexibility allows extensions of the Hough transform to novel domains such as object tracking and action recognition. Hough forests can be considered as task-adapted codebooks of local appearance that permit fast supervised training and quick matching at test time. They attain great detection accuracy since the entries of such codebooks are optimized to cast Hough votes with small variance, and since their effectiveness allows intense sampling of local image patches or video cuboids during detection. The efficacy of Hough forests for a set of computer vision tasks is authorized through experiments on a huge set of publicly accessible benchmark datasets and comparisons with the state-of-the-art.

There has been a rising interest in exploiting appropriate information in accumulation to local features to identify and localize various object categories in an image. A context model can rule out some improbable combinations or locations of objects and guide detectors to create a semantically coherent interpretation of a scene. However, the performance benefit of context models has been limited because most of the previous methods were tested on datasets with only a little object categories, in which most images hold one or two object categories. In this paper, ***Myung Jin Choi et al [5]*** introduce a novel dataset with imSages that include various illustrations of dissimilar object categories, and propose an efficient model that confines the contextual information among more than a hundred object categories utilizing a tree structure. Given model incorporates global image features, dependencies between object categories, and results of local detectors into one probabilistic framework. They demonstrate that provided context model improves object recognition performance and provides a coherent interpretation of a scene, which enables a reliable image querying system by multiple object categories. In addition, this model can be employed to scene understanding tasks that local detectors alone cannot resolve, such as identifying objects out of context or querying for the most classic and the least typical scenes in a dataset.

Successful state-of-the-art object recognition techniques from images have been based on powerful methods, such as sparse representation, in order to replace the also popular vector quantization (VQ) approach. Recently, sparse coding, which is characterized by representing a signal in a sparse space, has raised the bar on several object recognition benchmarks. However, one severe disadvantage of sparse space based methods is that parallel local features can be quantized into dissimilar visual words. ***Gabriel L. Oliveira et al [6]*** presents in this paper a new method, called Sparse Spatial Coding (SSC), which combines sparse coding dictionary learning, a spatial constraint coding stage and an online classification method to improve object recognition. An efficient new off-line classification algorithm is also presented. They overcome the problem of techniques which make use of sparse representation alone by creating the final representation with SSC and max pooling, offered for an online learning classifier. Experimental results achieved on the Caltech 101, Caltech 256, Corel 5000 and Corel 10000 databases, show that, to the best of knowledge, this approach supersedes in accuracy the best published results to date on the same databases. As an extension, they also show high performance results on the MIT-67 indoor scene detection dataset.

3. BACKGROUND

From the study of different object recognition models that is observed that there are a number of learning methods are available for object recognition. Some of them are works on the pose based techniques [10], part based, colour and texture based [12] approaches and similar methods. Therefore in order to demonstrate new object recognition technique human sentiment detection based object recognition technique is desired to develop in this proposed study.

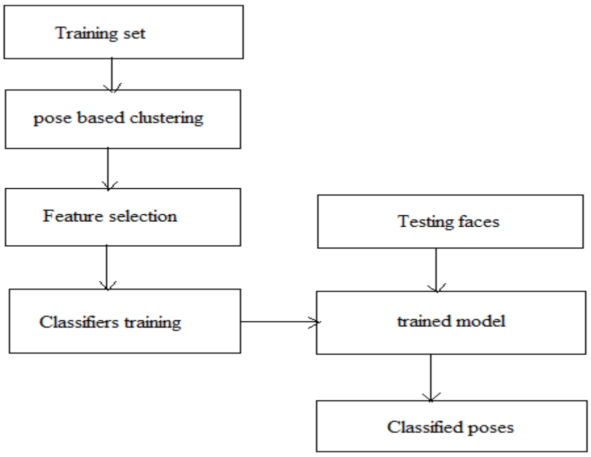

Fig 1 Proposed object recognition model

In order to understand the proposed object recognition model different concepts are utilized, which includes pose based learning, visual feature extraction and selection techniques, classification and recognition algorithms. a conceptual incorporated model and their subcomponents are given using figure 1. And their subcomponents can be discussed as:

3.1 Training Set

In machine learning and pattern recognition process the training set played and essential role. For accurate learning and recognition it is desired to have a data model specific training set. In other words if the data set contains the noise and outliers then the learning model trained with these noisy and invalid pattern which may misguide the final learning model.

The machine learning is classified according to the acceptance of their training set, the models which are learn with the attributes and their pre-defined class labels are known as supervised learning techniques. On the other hand models which learn only with the attributes and prepare their guidelines self are termed as un-supervised learning techniques.

In this proposed work for analysing the objects and their emotions a similar object faces are considered with a number of poses, a simple training set example is given using table 1.

Table 1 Training Set

	Laughing face
	Serious

	Thinking
	Normal

The table (1) demonstrate the training set example, which contains a set of images of same person with different emotions. Here the different faces are the attributes of learning model, and the emotions are target class which is desired to recognize. For efficient and effective learning first required to find prepare a training set with a number of objects and their different and similar poses.

3.2 Pose base Clustering

That is second phase of the proposed object recognition process model. In this phase first data is clustered or grouped according to their poses. Now a data is transformed and a set of similar poses are incorporated for targeting the single class. The table 2 demonstrate the transformed data for recognising the single emotions.

Table 2 Similar Class Object

Laughing faces

The object recognition process needs a significant amount of training data for effective learning. Therefore, a considerable amount of memory and time is consumed due to the size of visual example. In the colour images or examples the definition of single pixel is provided using three pixel values. Therefore with the pose based clustering it is required to decrease the quantity of data, for that purpose two processes are incorporated.

1. **Grey scale conversion:** In this step the colour images are converted into grey scaled images, thus the size of images is condensed, because is this type of images a single value is able to represent a pixel. For providing a single grey value $G_{i,j} = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{R_{i,j}+G_{i,j}+B_{i,j}}{3}$formula can be used.
2. Feature extraction: In first phase the quantity of data is reduced more and image features are computed. There are three key features are basically evaluated for extracting meaningful patterns.
    1. Colour: the colour feature of an instance data is represents the colour distribution in the target object. In multi-pose object detection colour properties are approximately similar for all over the object surface.
    2. Shape: the shape feature represents the outlines or edges of the image objects, therefore it significantly extract the meaningful pattern over data for object recognition process.
    3. And texture: the texture of image object denotes the image quality, and their internal pixel organization. Therefore the estimation of the texture feature is not much necessary.

Thus in order to define the objects in our model only shape feature is required to incorporate with learning method. In next section we find some appropriate method for shape feature extraction.

4. SHAPE FEATURE EXTRACTION METHODS

There are various kinds of algorithms are available that are promises to provide efficient edge detection during feature vector calculation. Some of them are listed in this section.

4.1 Canny Edge Detection

The principle of edge detection in common is to significantly decrease the quantity of data in an image, while preserving the structural properties to be used for further image processing. Numerous algorithms exists, and this worksheet aims on a specific one developed by John F. Canny (JFC) in 1986. [7, 9]

The algorithm runs in 5 separate steps:

1. **Smoothing:** Blurring of the image to remove noise.
2. **Finding gradients:** The edges should be marked where the gradients of the image has huge magnitudes.
3. **Non-maximum suppression:** Only local maxima should be marked as edges.
4. **Double thresholding:** Potential edges are determined by thresholding.

5. **Edge tracking by hysteresis:** Final edges are resolute by suppressing all edges that are not linked to a very certain (strong) edge.

4.2 Smoothing

It is inevitable that all images taken from a camera will include some quantity of noise. To stop that noise is mistaken for edges, noise must be decreased. Hence the image is first smoothed by applying a Gaussian filter. The kernel of Gaussian filter with a standard deviation σ = 1.4. The effect of smoothing the test image with this filter is shown in Figure 2.

4.3 Finding Gradients

The gradient magnitudes (also termed as the edge strengths) can then be determined as Euclidean distance evaluate by applying the law of Pythagoras.

$$|G| = \sqrt{G_x^2 + G_y^2}$$

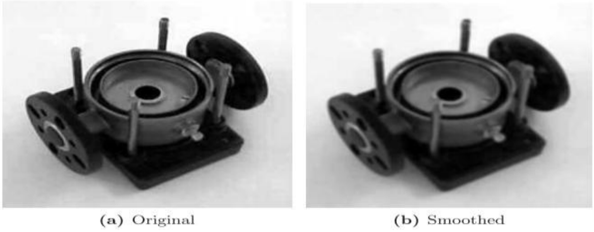

Figure 2 Smoothing effect on image

It is sometimes simplified by applying Manhattan distance measure to reduce the computational complexity.

$$|G| = |G_x| + |G_y|$$

Gx and Gy are the gradients in the x- and y-directions respectively.

The Euclidean distance measure has been applied to the test image. The computed edge strengths are compared to the smoothed image in Figure 3.

That an image of the gradient magnitudes often indicates the edges quite clearly, However, the edges are typically broad and thus doing not indicate exactly where the edges are. To make it feasible to determine this, the direction of the edges must be identified and stored as.

$$\theta = arcTan\left(\frac{|G_y|}{|G_x|}\right)$$

4.4 Non-maximum Suppression

The purpose of this step is to alter the "blurred" edges in the image of the gradient magnitudes to "sharp" edges. Principally this is done by conserving all local maxima in the gradient image, and deleting everything else. The algorithm is for every pixel in the gradient image:

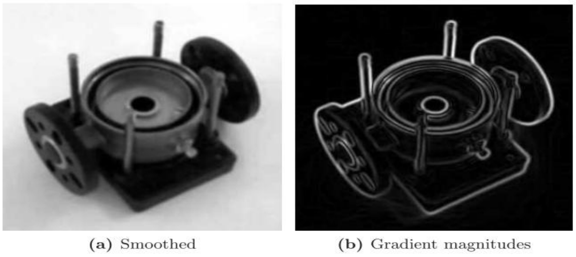

Figure 3 Gradient magnitudes of image

1. Round the gradient direction θ to nearest 45◦, corresponding to the use of an 8-connected neighbour-hood.
2. Evaluate the edge strength of the present pixel with the edge strength of the pixel in the positive and negative gradient direction. I.e. if the gradient direction is north (theta = 90◦), evaluate with the pixels to the north and south.
3. If the edge strength is largest of the current pixel; protect the value of the edgse strength. If not, suppress (i.e. remove) the value.

4.5 Double Thresholding

The edge-pixels residual after the non-maximum suppression step are (still) marked with their strength pixel-by-pixel. Many of these will probably be true edges in the image, but some maybe caused by noise or color variations for instance due to rough surfaces. The easy way to discern between these would be to use a threshold, so that only edges stronger that a definite value would be conserved. The Canny edge detection algorithm uses double thresholding. Edge pixels stronger than the high threshold are marked as strong; edge pixels weaker than the low threshold are concealed and edge pixels between the two thresholds are marked as weak.

4.6 Edge Tracking by Hysteresis

Strong edges are interpreted as "certain edges", and can instantly be included in the final edge image. Weak edges are integrated if and only if they are linked to strong edges. The logic is of course that noise and other minute variations are not likely to result in a strong edge (with appropriate adjustment of the threshold levels).

Thus strong edges will (almost) only be due to true edges in the original image. The weak edges can either be due to true edges or noise/color variations. The latter type will probably be distributed independently of edges on the whole image, and thus only a small quantity will be situated adjacent to strong edges. Weak edges due to true edges are much more liable to be linked directly to strong edges.

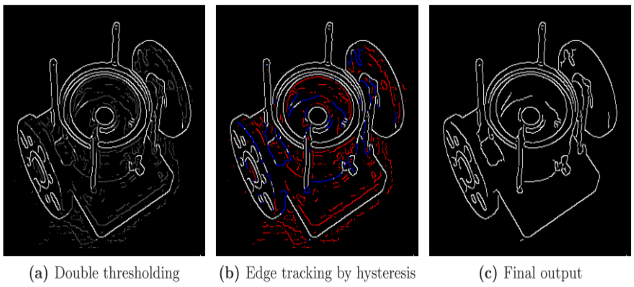

Figure 4 Blob Analysis

Edge tracking can be implemented by BLOB-analysis (Binary Large Object). The edge pixels are separated into linked BLOB's using 8-connected neighbourhood. BLOB's including at least one strong edge pixel is then preserved, while other BLOB's are suppressed. The effect of edge tracking on the test image is shown in Figure 4.

4.7 Sobel Operator

The operator consists of a pair of 3×3 convolution kernels as shown in given matrix [13]. One kernel is simply the other rotated by 90°

-1	0	+1
-2	0	+2
3	0	+3

+1	+2	+3
0	0	0
-1	-2	-3

Figure 5 Sobel Operator

These kernels are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid, one kernel for each of the two perpendicular orientations. The kernels can be applied separately to the input image, to produce separate measurements of the gradient component in each orientation (call these Gx and Gy). These can then be combined together to find the absolute magnitude of the gradient at each point and the orientation of that gradient. The gradient magnitude is given by:

$$|G| = \sqrt{G_x^2 + G_y^2}$$

Typically, an approximate magnitude is computed using:

$$|G| = |G_x| + |G_y|$$

which is much faster to compute.

The angle of orientation of the edge (relative to the pixel grid) giving rise to the spatial gradient is given by:

$$\theta = \arctan\left(\frac{G_y}{G_x}\right)$$

4.8 Robert's Cross Operator

The Roberts Cross operator performs an easy, rapid to calculate, 2-D spatial gradient measurement on an image. Pixel values at every point in the output symbolize the expected absolute magnitude of the spatial gradient of the input image at that point. The operator includes of a pair of 2×2 convolution kernels as shown in Figure 2.5 One kernel is simply the other rotated by 90°. This is very similar to the Sobel operator.

+1	0
0	-1

0	+1
-1	0

Figure 6 Robert operator

These kernels are considered to respond maximally to edges running at 45° to the pixel grid, one kernel for each of the two perpendicular orientations. The kernels can be applied separately to the input image, to create separate measurements of the gradient component in each orientation (call these Gx and Gy). These can then be mixed together to find the absolute magnitude of the gradient at every point and the orientation of that gradient. The gradient magnitude is given by:

$$|G| = \sqrt{G_x^2 + G_y^2}$$

Although typically, an approximate magnitude is computed using:

$$|G| = |G_x| + |G_y|$$

which is much faster to compute. The angle of orientation of the edge providing grow to the spatial gradient (relative to the pixel grid orientation) is given by:

$$\theta = \arctan\left(\frac{G_y}{G_x}\right) - \frac{3\pi}{4}$$

4.9 Prewitt's Operator

Prewitt operator is identical to the Sobel operator and is utilized for detecting vertical and horizontal edges in images.

-1	0	+1
-1	0	+1
-1	0	+1

+1	+1	+1
0	0	0
-1	-1	-1

Figure 7 Prewitt gradient edge detectors

5. CONCLUSION AND FUTURE WORK

The proposed work is intended to develop a pose based object recognition model. For demonstrating such kind of model various object recognition models are studied and based on their concepts, an emotion detection technique is desired to develop. Therefore an overview of complete system organization is discussed. The proposed system is a modular system, thus first two modules and their explanation is also reported in this paper. In addition of that, a brief review on edge detection technique is also provided, that may help on selecting the appropriate and efficient smooth edge detection technique.

Thus in near future with the selection of edge detection algorithm and extended design of proposed object recognition model is presented.

6. REFERENCES

[1] R. Lefort, R. Fablet and J.-M. Boucher, "Object recognition using proportion-based prior information: Application to fisheries acoustics", Pattern Recognition Letters January 2011, Volume 32, Issue 2, Pages 153-158

[2] Zhangzhang Si and Song-Chun Zhu, "Learning AND-OR Templates for Object Recognition and Detection", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 35, NO. 9, SEPTEMBER 2013

[3] Z. Si and S.-C. Zhu, "Learning Hybrid Image Templates (HIT) by Information Projection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 7, pp. 1354-1367, July 2012.

[4] Juergen Gall, Angela Yao, NimaRazavi, Luc Van Gool, and Victor Lempitsky, "Hough Forests for Object Detection, Tracking, and Action Recognition", IEEE Transactionson Pattern Analysis and machine intelligence, VOL. X, NO. X, JANUARY 2011

[5] Myung Jin Choi, Joseph J. Lim, Antonio Torralba, Alan S. Willsky, "A Tree-Based Context Model for Object Recognition", Computer Science and Artificial

Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2010-050 October 29, 2010

[6] Gabriel L. Oliveira, Erickson R. Nascimento, Antonio W. Vieira, Mario F. M. Campos, "Sparse Spatial Coding: A Novel Approach for Efficient and Accurate Object Recognition", 2012 IEEE International Conference onRobotics and Automation (ICRA)

[7] Canny Edge Detection, March 23, 2009

[8] ZhenhuaGuo,LeiZhang,David Zhang, "A Completed Modeling of Local Binary Pattern Operator for Texture Classification", IEEE transaction on image processing, 2010

[9] MasoudNosrati, RonakKarimi, Mehdi Hariri, "Detecting Circular Shapes From Areal Images Using Median Filter and CHT", World Applied Programming, Vol (2), Issue (1), . 49-54, January 2012

[10] Bangpeng Yao, Li Fei-Fei, "Modeling Mutual Context of Object and Human Posein Human-Object Interaction Activities", 2010 IEEE Conference onComputer Vision and Pattern Recognition (CVPR)

[11] Kevin Lai,Liefeng Bo,XiaofengRen, Dieter Fox,"A Scalable Tree-Based Approachfor Joint Object and Pose Recognition", Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

[12] ArjanGijsenij, and Theo Gevers, "Color Constancy UsingNatural Image Statistics and Scene Semantics", IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol . 33, No. 4, April 2011, pp 687

[13] WenshuoGao, Lei Yang, Xiaoguang Zhang, Huizhong Liu, "An Improved Sobel Edge Detection", 978-1-4244-5540-9/10/$26.00 ©2010 IEEE

[14] Alvaro Collet, Manuel Martinez, Siddhartha S. Srinivasa, "The MOPED framework: Object Recognition and Pose Estimation forManipulation", 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)