Object Recognition using Template Matching and Pre-trained convolutional neural network

Qaisar Abbas†

†College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

Summary
Template Based Object Recognition (TBOR) is very active research area in different fields for finding object in a video or image. Nowadays, it is very difficult for TBOR-system to classify multiple-objects due to use of conventional machine learning algorithms and higher computational complexity for manually tuned features. According to literature review, it was noticed that most of existing TBOR system were unable to focus on multiple objects. To overcome these problems, a new TBOR approach is developed in this paper for the recognition of multiple objects by combining both techniques such as template-based matching and a pre-trained convolutional neural network (CNN) model. In the proposed TBOR system, object images are first projected onto features space known as template space that best encodes the variation among known object of templates. The template space is then defined by Eigen faces, which are the eigenvectors of set of objects. Afterwards, principal component analysis (PCA) method is applied to find the approximate aspects of objects, which are important for identification. At last, object recognition step is performed by combing template based PCA and pre-train CNN methods. A template-based matching technique was fully automatically implemented in this study through PCA analysis to initially recognize the object using correlation and phase angle methods. The recognition results are further enhanced by pre-train CNN model. Experimental results indicate that the proposed system is outperformed compared to state-of-the-art template-matching algorithms in terms of accuracy.

Key words:
Computer vision, object recognition, template matching, deep learning, convolutional neural network, principle Component Analysis

1. Introduction

Recognition of objects from a video scene is a difficult and important step. In practice, humans are more expert to classify different objects from an environment or from live video stream regardless of the fact that the image of the objects may have different scale, orientation and sizes even when they are translated or rotated [1]. In domain of computer vision, this task is very challenging. However, several authors developed many systems to recognize the objects from scenes. In fact, those systems were unable to achieved higher accuracy due to different shapes of objects. In fact, the object recognition is a particular field of learning. Currently, many authors developed systems for classification of multiple objects from still or moving images using template-based or deep-learning algorithms.
Whatever techniques used in the modern object classification systems, there is always learning step required and system has to train on a set of certain sample images. When learning system was trained then it can help to recognize new objects class. In general, object classification task is particularly related to computer vision, which is divided into stages such as low-level vision and high-level vision. In case of low-level vision, objects were isolated from the image regions based on features. Whereas high-level vision tasks are used extract objects from a reference scene based on correlation measure. In general behavior, a correlation measure is a degree of two variables corresponding pixel values in two images known as template and source. Template Matching is a state-of-the-art method to classification objects.
Several methods have been developed to classify objects using template-based matching (TBOR) algorithms [2-5]. Those TBOR systems have many different applications in practice. To measure similarity among objects in a scene, there are famous techniques used in the past such as normalized cross correlation (NCC) and square root of Sum of square differences (SSD) [6,7]. Compare to scene analysis TBOR approaches, many other authors used sum of absolute differences (SAD) and sequence similarity detection algorithms (SSDA) [8] to identify different pattern from digital images.
In this paper, a template base object Recognition (TBOR) system studies the work done in the field of Template Matching and introduces a new approach to the problem of Recognition of Multiple Objects. Also, this thesis focuses on the study of how templates are matched with input image and how the objects are recognized by using template matching. The problem statement can be described as follows. Given an image of object, it can be human being, bird, chair or some other object, compare it with templates (models) [9] stored in the data set. Data set is database of object templates. And at the end, report what it is? It shows the percentage of matching with each type of template stored in data set. If a match exists it displays the name of that object.

Manuscript received August 5, 2020
Manuscript revised August 20, 2020

The image taken is color image. This input image can have more than one object to be recognized. That is, it

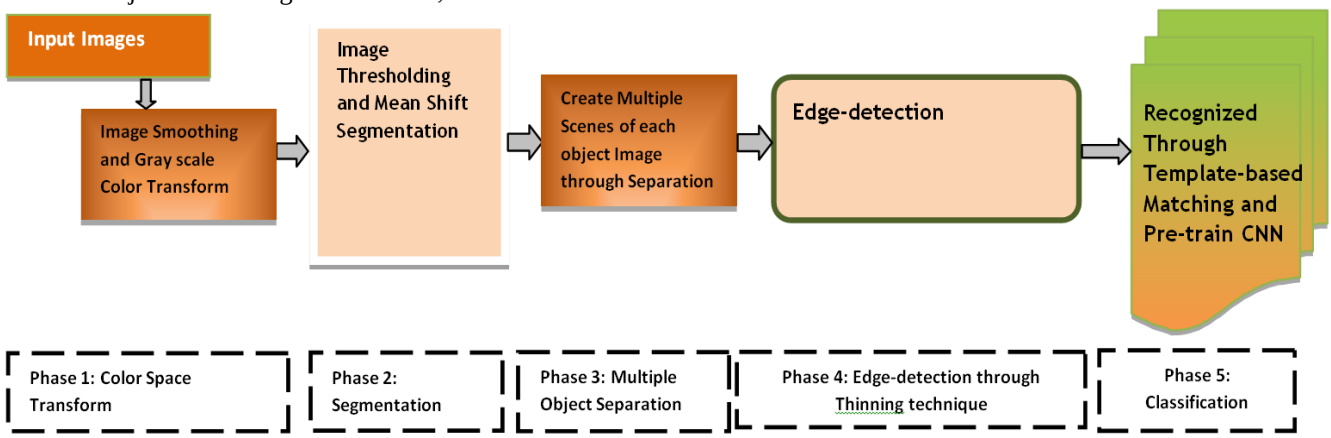


Fig. 1 Template-based matching system for detect and track multiple objects using pre-train CNN model.

performs matching with multiple objects, and recognizes that object if it exists. I have used templates of human beings, birds and chairs in this paper. That is data set of images contains templates of human beings, birds and chair. This data set can be extended to as many objects or templates as you want. To do all this, I have divided this paper into three basic modules; Image processing, Image segmentation and Recognition using Principal Component Analysis and CNN model [10].

In image processing module, image smoothing and converting image to gray scale has been done. In second module or in Image Segmentation module, image thresholding step was performed. It is first step towards image segmentation. Image thresholding separates object pixels from background pixels. Then I have used mean-shift approach for object segmentation. The third module of this paper is to match segmented object with templates stored in data set. For this, I have used Principal Component Analysis PCA and CNN model. Many algorithms have been proposed for template matching. But most of them work with single object. So, for the recognition of multiple objects, I have implemented Principal Component Analysis (PCA) with neural network.

2. Related Work

Several automatic template-based matching algorithms were developed in the past systems. For example in ref. [11], the authors recognized basic objects based on shape or structure. To make difference, the authors proposed a template-based model without using pre-processing steps. Without using image processing techniques, learning of template is very difficult to get high accuracy of classification. However in that study, the authors achieved better results compare to state-of-the-art algorithms. A 3D model based different approached was developed in [12] to detect and track 3D objects through template-based detection method. The authors claimed that they improved a framework by 13% compared to the state-of-the-art system. They noticed that there proposed systems were cap abled to work in real-time application. In that approach, the authors did not perform any image processing step to extract visual features.

However in [13], an improved system was presented based on video-based feature descriptor and template-based classifier. In that paper, the authors performed feature selection step that can be used to real-time track object. This study is different compared to state-of-the-art systems. Also in [14], the authors developed an efficient system to capture the different modalities through template-based matching concept. They integrated dense depth map and complementary object information techniques to complete this study.

In [15], an effective shape-based template-based approach was presented to segment both local and global parts of the objects that were appeared in images. They applied this approach on detecting humans and their poses through template-based learning methods. In that study, they used location-based encoding schedule to extract features. Those features are trained by a train kernel-SVM classifier to distinguish among human and other object patterns. They tested this approach to detect multiple occluded human. This approach demonstrates that they achieved higher accuracy to detect multiple human from image even if they were occluded from other objects. Whereas in [16], authors utilized a SIFT algorithm to recognize face based on multi-scale local features. In that system, they used a single template per subject to make difference in patterns.

In [17], authors presented a new template-based system to recognize speed-limit-sign through GPU-based processing in an embedded manner. They used a pre-processing step to do contrast enhancement by composite filters. They acquired higher 90% accuracy when tested on 45 minutes of road-video compared to other systems. A real-time 3D object detection method was presented in [18] through template matching. They utilized many templates in real-time with a dense depth sensor to recognize objects. To use templates, the authors used in [19] to recognize vehicle license plate by using a hybrid classifier that recognize characters on it. Moreover, a hybrid classifier was developed through SVM. A decision tree was used to recognize the plate characters. The authors reported that they achieved 96% performance value.

In [20], template-based matching is also applied on assessment of dietary products to monitor energy and nutrient intakes. To estimate food portion size, they used built-in mobile camera to take picture. As a result, it is very much difficult to determine nutrient content of foods through image-based dietary assessment. In that study, the authors segment food images through template-food image shape. Also in [21], the authors used template-based approach to detect and track objects from serious of images. They also utilized SIFT features to generate the feature points of an object template. After this, a feature matching method was applied to classify objects in images. Whereas in [22], a new methods was developed to detect gesture from motion sequences based on template-based matching technique. Authors claimed that the suggested method did not require any prior knowledge about foreground or background.

In [23], a template-based method was applied on rubber tree leaflets for leaflet clone classification. To detect overlap among rubber tree, a template-based approach was adopted to extract key point based features. Also in [24], the authors utilized deep learning algorithm along with template-based matching approach for face recognition. They used multilayer conventional neural network (CNN) model to extract the features and then template-based approach was performed to recognize human face.

3. Methodology

The systematic flow diagram of proposed template-based matching (TBOR) system is visually represented in Fig.1. In this paper, TBOR system consists of image processing steps to enhance the digital image and then used template-based matching and pre-trained CNN model to recognize the multiple objects. The detail of each development step was presented in the subsequent paragraphs. However first, a methodological background on template-based matching algorithms is presented in the sub-section A. After that, a proposed system was described in detail.

3.1 Template-based matching Algorithms

The interest in this matching problem or template matching field is not new. In the 1940's, at the Massachusetts Institute of Technology (MIT), the development of a new pattern matching algorithm began. The idea was to simulate the human brain's ability to learn and recognize patterns. First solutions for image matching have also been suggested in the late fifties especially by “Hobrough” in 1959. Since then a steady increase in the interest for image matching has occurred. There are many different methods of recognition by template matching. Matching methods differ according to the type of features used. For example, edge pixels, intensity image patches etc. Correlation techniques are widely used for template matching. These are following as mentioned below.

1) Correlation
2) Cross-Correlation
3) Normalized Cross-correlation

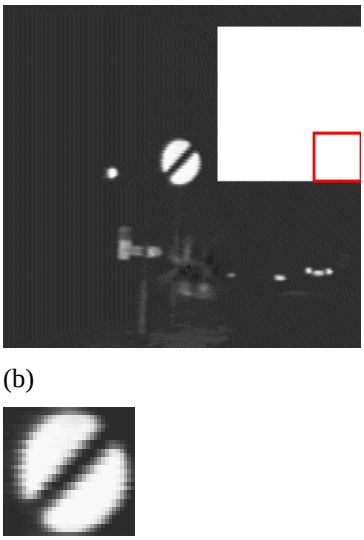
First, correlation is the simplest method of template matching. It is a traditional matching algorithm based on the correlation between portions of the image and a template. In this algorithm, mathematical shifting by rows and columns is used around the new image until a correlation peak is found. The higher the intensity of the peak found, the better the match to the original pattern. The template is typically an image of the object to be found, and correlation is used to find the location within the image that best matches the template. This technique requires that a template exist. Correlation techniques work well when scale and orientation (direction) are known must have a template, otherwise wrong evaluation of result may occur. Brightness sensitivity.

The following examples deal with the two major limitation of the correlation algorithm.

a. Brightness Sensitivity
b. Evaluation of the result

Correlation method of template matching is brightness sensitive. As, it find the maximum measure of similarity between input image and template. It may give wrong results when more bright areas exist in input image.

$$d_{f,t}{}^2(u,v) = \sum_{x,y}\left[f^2(x,y) - 2f(x,y)t(x-u,y-v) + t^2(x-u,y-v)\right]$$
$$= \sum_{x,y} f^2(x,y) + \sum_{x,y} t^2(x-u,y-v) - 2\sum_{x,y} f(x,y)t(x-u,y-v) \quad (1)$$

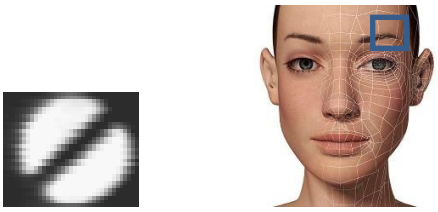


(b)

(a)

Fig. 2 Correlation (a) Template (b) location error due to brightness White square or brightness in the image leads the algorithm to a false result; the best match is not at the "good" place.

For template matching with correlation, it is must that that input image should contain template. If the template do not exist in input image, then it point out that template somewhere else in input image.



(a) (b)

Fig. 3 Correlation Problem (a) Template (b) input image

There is no template in the input image but after the correlation, a peak and the algorithm are continued and find a template where there is nothing. Similarly to find the location of equivalent patches of the image, the cross-correlation algorithm were designed. Those algorithms were find the similarity based on gray levels of pixels. Afterwards, the reference image (template) was searched from the image based on the reference point. Moreover, the reference image is moved in the search image to achieve this task based on maximum similarity of gray-levels at that position. So, a similarity value is considered at teach position of the reference image in the examine image. The use of cross-correlation for template matching is motivated by the distance measure (squared Euclidean distance). It is based on the Sum of Squared Difference (SSD) and derived as:

$$d_{f,t}{}^{2}(u,v)=\sum_{x,y}[f(x,y)-t(x-u,y-v)]^{2} \quad (2)$$

Where f is the image and the sum is over x,y under the window containing the feature t positioned at u,v. If the template and the image align exactly then the distance between them is zero ($d^2=0$), otherwise we will have $d^2>0$ and a large number means a poor match. Hence d^2 can be considered as a "mismatch" measure or distance measure. Some improvement over the previous equation can lead to the following alternative at the SSD. By simplifying previous equation, the expansion of d^2 in Eq. (2) can be rewritten as in Eq. (3).

The term $\Sigma\ t^2(x-u, y-v)$ is constant. If the term $\Sigma\ f^2(x, y)$ is approximately constant then the remaining cross-

$$c(u,v)=\sum_{x,y} f(x,y)t(x-u,y-v) \quad (3)$$

Correlation term is a measure of the similarity between the image and the feature. This term is maximum when the portions of the image "under" the template are the same. And if this term is maximum then the distance measure is minimum (means a good match of template and image). This new equation is used in the cross-correlation algorithm, as the best match is where the result of the correlation is maximum. The following example shows the result of a template matching using the Cross- Correlation algorithm. Square in input image indicates where the algorithm finds the template.

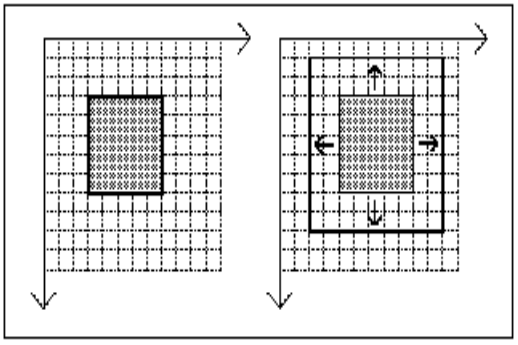
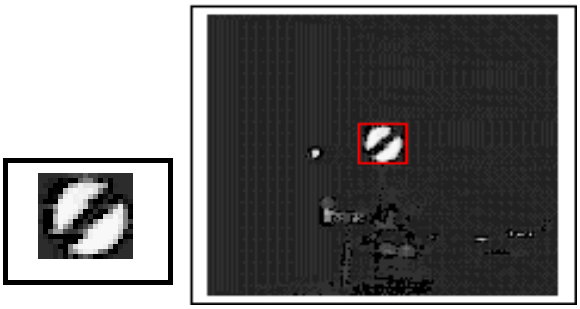


Fig. 4 A visual example of principle of cross correlation technique.



(a) (b)

Fig. 5 Cross-Correlation Template input and result image

This correlation technique has several disadvantages:

- Computational expensive (but special chips available)
- Scale/rotation sensitive. It is important in most applications to know when a match has a significant uncertainty, then the possibility exists, that a qualitatively incorrect position has been found.

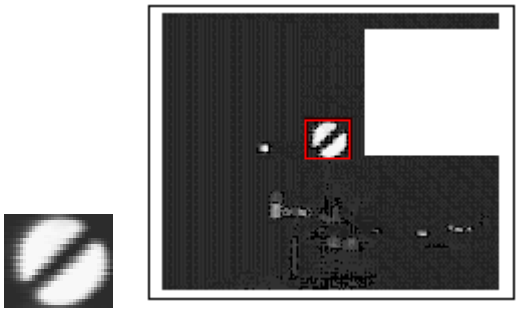
This Equation is derived under the supposition that the image energy $\Sigma f^2(x,y)$ is approximately persistent. As a result, if this energy factor varies with position then it will be difficult to find a perfect match and algorithm will definitely fail. The reason is that the correlation between the features and a bright spot is greater than the correlation between them. The range of c(u,v) is totally dependent on the size of the feature, which is presented in an image. The normalized Cross-correlation is mainly based on the correlation algorithm; in fact it is just an improvement that solves two problems of the correlation algorithm.

- Evaluation of the result
- Brightness sensitivity

$$\rho(u,v) = \frac{\sum_{x,y}\left[f(x,y) - \overline{f}_{u,v}\right]\left[t(x-u,y-v) - \bar{t}\right]}{\sqrt{\left(\sum_{x,y}\left[f(x,y) - \overline{f}_{u,v}\right]^2 \sum_{x,y}\left[t(x-u,y-v) - \bar{t}\right]^2\right)}} \quad (4)$$

Where $\bar{t}$ is the mean of the template and $\bar{f}$ is the mean of f(x,y) in the region under the template. This equation is known as "Normalized cross-correlation" or "Normalized grayscale correlation". The correlation coefficient lies between –1 and 1. That is $-1 \leq \gamma(u, v) \leq 1$. When $-1 \leq \gamma(u, v) \leq 0$, the template and the region of the image under the template at position u,v is totally uncorrelated. When $\gamma(u, v)$ is equal to 1, it means that we have a perfect match at position u,v.

The example below show that the NCC (Normalized Cross-Correlation) could succeed where the standard correlation fail. But it's not the only improvement of the NCC over the standard correlation; this algorithm let us know when a result is a really good match. As the coefficient, it returns is normalized between –1 and 1; if the result is under 0 there is no need to go further because the template and the image at the point are uncorrelated. If the result is 1, it is the perfect match but if it is 0.5 we can say it has only 50 % of accuracy.



(a) (b)

Fig. 6 Normalized Cross-Correlations (a) Template (b) Success on brightness

The NCC solves some of the limitations of the correlation algorithm but not all of them. Especially, the NCC is (like the standard correlation) unable to locate the template when there is a rotation or a scale modification. The first and major drawback of almost all of template matching algorithm is that they work with only one template, not with multiple templates. Because they simply roll the template over the input image and compute the position of maximum similarity. And note the position where template is maximum similar to the template. If two or more templates are passed to these algorithms, they will find the position of maximum similarity of every passed template, without being care that this template (object) exists or not in image. Almost all of matching algorithm do not work with multiple objects. They are just matching algorithm, they cannot be used for recognition. Because they do not know what is the input, and what is the output. They simply compute position where template maximum matches to image. That's why I have used Neural Network using Principal Component Analysis (PCA) in this study. PCA is a tested algorithm for face recognition. And I think, if an algorithm can be used for face recognition, then it can be used everywhere.

3.2 Proposed Template-based matching Algorithms

This paper is related to object recognition that performs matching with multiple templates. In this study, I have tried

to overcome the draw backs of Template Matching Algorithm. For this I have used a stepwise approach for recognition. This paper is divided into different modules such as pre-processing, initial segmentation, segmentation Refinement and object classification.

When an image is taken as input image, it can be color image or it may not be smooth. Then first image smoothing is performed to smooth color image, and then color image is converted to grayscale image. These two steps are known as pre-processing of any image processing technique. That is, in any paper of image processing, these two steps are performed for good results. Since the image processing is the field in which human eye is to be simulate. In first step I have performed Gaussian Smoothing, then I have converted RGB (color) image to grayscale. Since many processing techniques, such as histogram, segmentation work on the intensity component of an image only. So, it becomes necessary to convert an RGB image into a gray scale image. For many image processing and computer vision tasks, object segmentation is an important basis. The goal of segmentation is to separate the object's pixel within an image from the background pixels. So, in segmentation, objects are separated from background. I have divided image Segmentation into following steps such as dynamic thresholding, mean-shift segmentation and edge detection.

Thresholding is an image segmentation process, based on converting a grey-scale image into a binary image by re-assigning pixel grey levels to only two values (0 and 1). Regions of the binary image are separated based on whether pixel values in the grey-scale image were above or below a chosen intensity level (also called Threshold value). Thresholding technique separates an 'object' from the background based upon the grey-level histogram of an image. If the grey-level values of an object within an image are quite different than the background grey-level value, then finding the optimum threshold value to threshold the image is quite easy. Threshold value is a point in the histogram of grey-values, which separates an object from background. In order to binaries an image, we have to find a suitable threshold to separate the image pixels into two classes of pixel value 0 and 1 (binary image). This technique is also called automatically picking a threshold value. In this method no fix threshold value is used. Instead, each time dynamically threshold value is calculated. This method is done by automatically picking a threshold value from a grayscale histogram.

Using this technique, threshold value can be dynamically calculated. In this technique a Histogram of grey Image is calculated. This Histogram has two high peaks, one represents object pixel's highest intensity and other represents background pixel's highest intensity. Using this histogram, threshold value can be dynamically calculated by using mode-method of Histogram.

If an image consists of objects that clearly differ from the background, the resulting histogram is bi-modal. Bi-modal Histogram has two highest peaks. One peak represents object and other represents background. To separate object pixel from background pixel, a threshold value is created between these two peaks. The purpose of doing this is to clearly identify object pixel from background pixel.

On x-axis possible grey level values (0-255) are shown, and on y-axis frequencies according to grey value have shown. Bi-modal histogram threshold detection algorithms usually find the highest peaks in the grayscale histogram, and detect the threshold as a minimum point between them. This technique is called the mode method. In this paper, I have used "mode method" to calculate Dynamic threshold point.

Thresholding essentially involves turning a color or grayscale image into a 1-bit binary image. This is done by allocating every pixel in the image either black or white, depending on their value. The threshold value is used to decide whether any given pixel is to be black or white.

Thresholding is used in image processing to separate an object's pixel from the background pixels. Thresholding converts a multi-grey level image into a binary image containing two grey-level values. The threshold operation is defined as:

$$g(x,y) = G0 \text{ if } f(x,y) > T$$

and

$$g(x,y) = G1 \text{ if } f(x,y) \leqslant T$$

Where f(x,y) is original image, g(x,y) is the binaries image, T is the threshold value, G0 is the object grey-level value after the thresholding operation, and G1 is the background grey-level values after thresholding operation.

3.2.1 Initial Segmentation

After converting image to Binary image, I have done Mean-Shift segmentation. For this, I have used mean-shift algorithm of segmentation. The mean-shift algorithm presented here was used by Comaniciu and Ramesh (2002). By Thresholding, the generated image is not uniform. Although Thresholding separates the background pixels from object pixels, but the separated objects are not clear to understand for further processing on these objects (or objects pixels). Since image Thresholding converts grey image into binary image which has only black and white pixels. So, the Threshold image may have very small areas of black pixels (object pixels). These pixels do not form an object. And also the Threshold image may have small area with white pixels within objects (i.e., objects can have gaps of white pixels in them). Image segmentation using mean-shift can be used to fix these problems. So, to remove small areas of black pixels that do not form objects, and to fill gaps within objects, I have used Image Segmentation using mean shift.

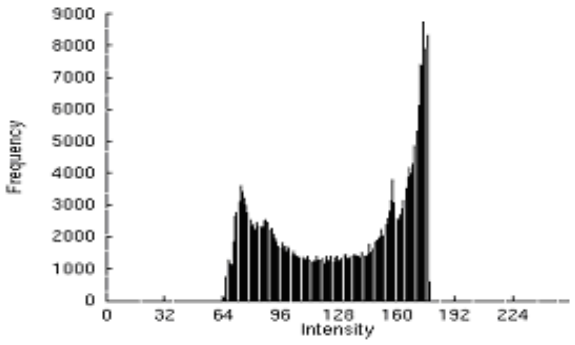


Fig. 7 An example of Bi-modal Histogram

The Mean-Shift based segmentation technique was introduced in 2002 by Comaniciu and Ramesh and has become widely-used in the vision community. It is one of many techniques under the heading of "feature space analysis".

The mean shift technique is comprised of two basic steps:

i. A mean shift filtering of the original image data (in feature space),
ii. And, a subsequent clustering of the filtered data points.

Mean shift filtering produces segmentations that correspond well to human perception. This algorithm is quite sensitive to its parameters. The mean shift filtering stage has two parameters corresponding to the bandwidths h (Radii of the kernel). The mean shift filtering step consists of finding the modes of the points and associating with them any points in the image space. This is to keep track of very small areas of black pixels, and also to keep track of the gaps in object pixels. For this the multivariate Kernel Density is used. After mean shift filtering, each data point in the feature space has been replaced by its corresponding mode. Some points may have collapsed to the same mode, but many have not despite the fact that they may be less than one kernel radius apart. It this step of mean-shift, small portion of black pixels are removed and gaps in object pixels are filled.

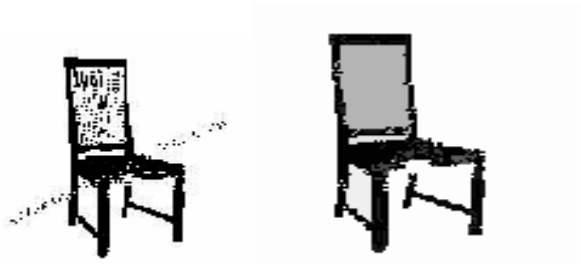


Fig. 8 An example of Mean-Shift algorithm

3.2.2. Refinement of Segmentation

After performing initial segmentation of objects from image, the pre-trained deep-learning model was used. This deep-learning model has been trained on humans and different objects datasets such as cars, chairs and vegetables. To refine the segmentation results, this convolutional neural network (CNN) model [16] is used to extract pixel-level features by using multi-layer architecture. This CNN model is having many applications in practice and it has been currently deployed for many object recognition tasks.

3.2.3. Object Classification

For recognition of object, a PCA is initially used by using neural network for matching. Neural Network and PCA techniques are best candidate for initial segmentation. To perform initial detection, neuron learning step is performed to update the weight of a particular neuron. This step is known as training processing where network is performed the verification.

This method tries to find the principal components in the distribution of features across images, also called the "eigenvectors". These eigenvectors can be considered as a set of features, which together characterize the variation between images. Each of the images in the training set can be represented exactly as linear combination of the eigenvectors. Since this linear combination is very large, so, by selecting some eigenvectors that have the largest eigenvalues, the most significant variation within the image set can be represented.

'Eigenvectors' can be described as a set of features, which together characterize the variation between template images (of objects). An Eigenvector is a vector (it must be non-zero vector). An 'Eigenvalue' of a square matrix is a scalar that is usually represented by the Greek letter λ.

In Mathematics, a number is called an eigenvalue of a matrix if there exists a non-zero vector such that the matrix times the vector, is equal to the same vector multiplied by the eigenvalues. Those vectors are then called the "eigenvector" associated with the "eigenvalue".

This relation between 'eigenvalue' and 'eigenvector' can be described by this equation:

For a given square Matrix 'A'

$Ax = \lambda x$ (suppose 'x' is an Eigenvector)

All Eigenvectors and Eigenvectors must satisfy this equation.

Each of the images in the training set can be represented exactly as linear combination of the eigenvectors. Linear combination is a Mathematical term that can be described as: suppose matrix 'A' has a set of 'n' linearly independent Eigenvectors x1 , x2 ,……, xn, corresponding to Eigenvalues λ1, λ2 ,……, λn (here λ1, λ2 ,……, λn need not to be distinct). Then $\lambda 1x1 + \lambda 2x2 + \lambda 3x3 + \ldots + \lambda nxn$

represent linear combination of eigenvectors. It can also be written as Axi = λixi for i=1, 2… n

Principal component analysis is used to find the aspects of objects which are important for identification. Eigenvectors are calculated from the initial data set. New objects are projected onto the space expanded by eigenvectors and represented by weighted sum of the eigenvectors. These weights are used to identify the objects. Principal Component Analysis (PCA) is a stepwise algorithm for recognition. These steps are:

Suppose we have to recognize the image 'I'. To do this, we have to follow these steps.

1. Subtract average template from input image I.
   $I = I - \psi$
2. Compute its projection onto the template-space
   $\Omega = U^{\top} I$
3. Compute the distance in the template space between the input image and all known template.
   $\varepsilon i^2 = \| \Omega - \Omega i \|^2$ for i = 1, 2, …… , M
4. Reconstruct the image from eigenvector, name it R
   $R = U\Omega$
5. Compute the difference between the image I and its reconstruction R
   $\xi^2 = \| I - R \|^2$

Where, ξ denotes difference between input image 'I' and its Reconstruction 'R'

On the basis of this difference we can recognize the input template. This difference can be divided into three categories:

If $\xi \geq \theta$, then its not any type of stored template.

If $\xi < \theta$ and $\varepsilon i \geq \theta$ (i = I ,2,….,M), then its new image.

Does not match to any stored template)

If $\xi < \theta$ and min { εi }$< \theta$, then its a known image (template exists in data set).

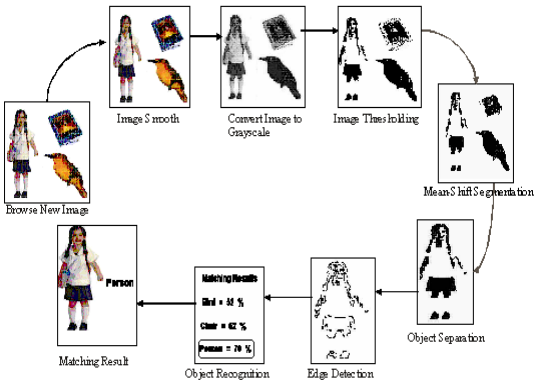


Fig. 9 Experimental results of proposed template-based matching algorithm

4. Experimental Results and Discussion

This Template Based Object Recognition (TBOR) system is implemented on Intel core I7 CPU, 16 GB of RAM, Windows operating system and python language. To test and compare the results with state-of-the-art TBOR systems, statistical measures were used as described in the sub-sequent sections. For the testing purpose of Template Base Object Recognition, here I will test it with different images (having more than one object) to recognize objects. The purpose of doing this is to check how much the software is accurate.

TBOR system has implemented through Principal Component Analysis (PCA) with Neural Network for object recognition. PCA is a tested algorithm of matching and it gives up to 90% of accurate matching results. Since it gives very accurate matching results, then it can also be used for recognition of objects with accurate results. That's why I have used PCA algorithm for recognition. I have tested it with different images. These input images have more than one object and they also have unknown objects (objects whose templates do not exists in data set). This study has shown up to 88 % of accurate matching results during testing. I have shown matching percentage of object with each template stored in data set. When the object don not exists in data set (template do not exists), then it displays it as "unknown object". In Figure 5.1, it is shown that an image of person is selected (form the input image) as the object to be recognized.

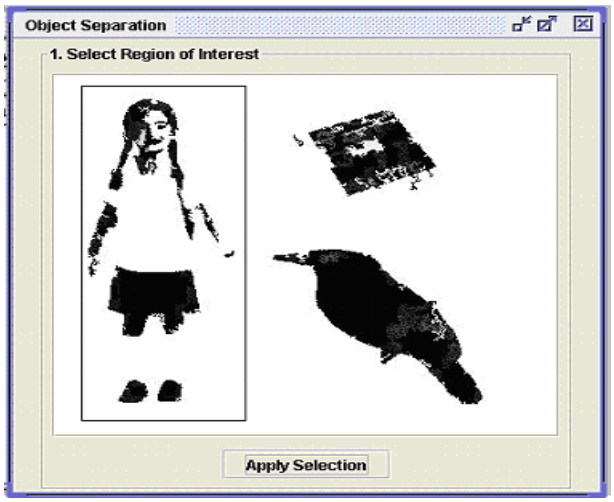


Fig. 10 An example of selecting Image of person

I have implemented templates of human being, bird and chair in my data set. Before recognition, neural network is trained for each template in data set. For the object recognition, a selected object (from input image) is matched with each template stored in data set. And the

percentage of matching of this selected object with all templates (in data set) is displayed in a frame. The template (in data set) with highest matched result is considered as the final result of recognition.

Similarly, 16 templates of 'human being' and 'Birds' are also stored in data set. This software dynamically accesses these templates and performs template matching with all of these templates. Since, it dynamically accesses these templates. So, it can be extended to as many objects (templates) as you want. The above Figure is showing the percentage of matching with chair, bird and person. Since the percentage of person is highest, so the final result 'Person' is displayed in "Result" frame. In Figure 5.4 a chair is to be send to neural network for recognition. A first improvement that we can do in this paper is automatic detection of multiple object s from input image. In this paper, I have used object cropping for object separation purpose. That is, user has to crop a desired area of image for recognition. And that matching is performed on this cropped area. Automatic detection can also be done for detection of objects. By using this, it will automatically detect objects from input image, and will perform template matching on these automatically detected areas. So we can consider different algorithms also. One algorithm that can be used is K. Sung and Poggio's "Example-based learning for detection". Object Recognition algorithm can be improved for better results of matching. I have implemented Turk and Pentland's Principal Component Analysis for the recognition of object. This algorithm was used by Turk and Pentland for the face recognition.

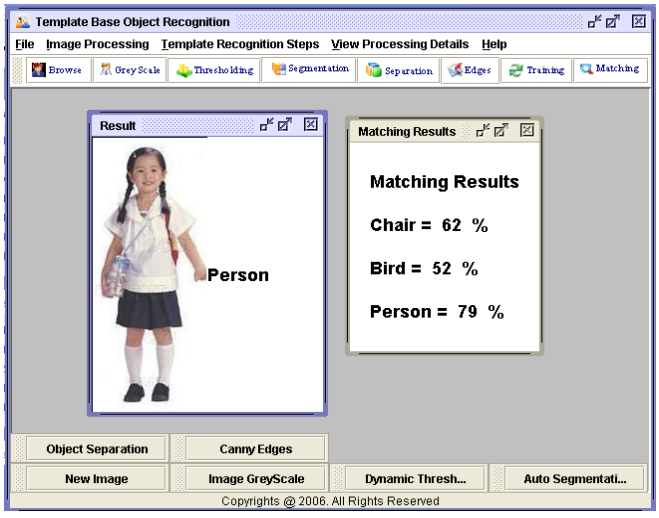


Fig. 12 A visual example of proposed system

Although in this paper, effective results are giving accurate matching result but nothing is perfect, there is always necessary need for improvement. So it can also be improved for more good results. The Principal Component Analysis PCA was the first working facial recognition system. In this study, an image is passed with multiple objects, but if the objects in image are very close, they effect on the accuracy of result. So, as an improvement some other recognition algorithm can be used than can provide much accurate results in case of very close objects in input image.

As another enhancement is that, it can be extended to work with video for object recognition. I have implemented this software with images having multiple objects, and this software recognize these objects. In future, it can be implemented with videos. Then it will be able to recognize object from real scene, and it will be able to use in robotics system. In robotics, a decision is made on the basis of recognition of object. So, after video implementation to this paper, it will be able to make real time decision where to fire and where not to fire.

To perform comparisons with state-of-the-art TBOR systems, the accuracy statistical measure is used on 4000 human and 25000 bird images. This dataset contains different size of images. To perform experiments, all those images are resized to (600x800) pixel resolutions. This dataset is acquired from different online sources. In this paper, the template-based recognition system was used from study developed in [13], [14] and [15]. Table 1 shows the experimental results in terms of accuracy and average time used to search an object from image. The accuracy in table 1 indicates that the percentage of the total number of correct detections from 100 repeated experiments. Moreover, the time column in the table shows that the time used in units of seconds to search an object.

To perform comparisons with [13], a template-based classifier was used to perform feature selection step used to real-time track object. Also in [14], an efficient system was used to capture the different modalities through template-based matching concept. The integrated dense depth map and complementary object information techniques were used to compare with TBOR system. A different study in [15] was used to compare with an effective shape-based template-based approach to segment both local and global parts of the objects that were appeared in images. This approach was applied on detecting humans and their poses through template-based learning methods. In that study, they used location-based encoding schedule to extract features. Those features are trained by a train kernel-SVM classifier to distinguish among human and other object patterns.

Table 1: State-of-the-art Comparison on 4000 human and 25000 of bird's dataset by using proposed TBOR with other three systems

No	*Methodology*	*Accuracy*	*Time (seconds)*
1	Template-based classifier[13]	78%	200s

2	Template-Features [14]	81%	250s
3	Shape-based-template [15]	84%	300s
4	Proposed- TBOR	92%	150s

On 4000 human and 25000 of bird images are used to perform experimental comparisons using detection accuracy measure. The table 1 shows the comparisons results and it indicates that the proposed TBOR system outperformed compare to other three approaches. This proposed Template Based Object Recognition (TBOR) is very effective to classify multiple-objects. However, other systems were mostly failed due to use of conventional machine learning algorithms and higher computational complexity for manually tuned features. According to literature review, it was noticed that most of existing TBOR system were unable to focus on multiple objects. To overcome these problems, a new TBOR approach is developed in this paper for the recognition of multiple objects by combining both techniques such as template-based matching and a pre-trained convolutional neural network (CNN) model. In the proposed TBOR system, object images are first projected onto features space known as template space that best encodes the variation among known object of templates. The template space is then defined by Eigen faces, which are the eigenvectors of set of objects. Afterwards, principal component analysis (PCA) method is applied to find the approximate aspects of objects, which are important for identification. At last, object recognition step is performed by combing template based PCA and pre-train CNN methods. A template-based matching technique was fully automatically implemented in this study through PCA analysis to initially recognize the object using correlation and phase angle methods. The recognition results are further enhanced by pre-train CNN model. Experimental results indicate that the proposed system is outperformed compared to state-of-the-art template-matching algorithms in terms of accuracy.

5. Conclusions

This paper explores the utilization of deep learning architectures in the hot topic of visual attention or saliency prediction. Two main variants of deep learning architectures are mostly employed in the state-of-the-art visual attention field such as convolutional neural network (CNN) and recurrent neural network (RNN) models. This paper is firstly explained the theory behind these two deep learning models. A first improvement that we can do in this paper is to develop an automatic detection of multiple objects from input image or video sequence through template-based matching and pre-trained CNN multilayer model. In this paper, object cropping for object separation purpose was utilized. That is, user has to crop a desired area of image for recognition and that matching is performed on this cropped area. Automatic detection can also be done for detection of objects. By using this, it automatically detects objects from input image, and then performs template-based matching on these automatically detected areas. So, a different algorithm is developed compared to state-of-the-art systems. If one algorithm that can be combined with this proposed system such as one developed by K.Sung and Poggio's "Example-based learning for detection". Then results of object Recognition algorithm can be improved for better results of matching.

Acknowledgment

The authors would like to express their cordial thanks to the department of Research and Development (R&D) of IMAM, university for research grant no: 360915.

References

[1] Zhang, W., Yao, T., Zhu, S., & Saddik, A. E. (2019). Deep Learning–Based Multimedia Analytics: A Review. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1s), 2.

[2] Wang, R., Liang, Y., Xu, J. W., & He, Z. H. (2019). Cascading classifier with discriminative multi-features for a specific 3D object real-time detection. The Visual Computer, 35(3), 399-414.

[3] Li, G., & Zhang, C. (2019). Automatic detection technology of sports athletes based on image recognition technology. EURASIP Journal on Image and Video Processing, 2019(1), 15.

[4] Prakash, S., Jayaraman, U., Gupta, P.: A skin-color and template based technique for automatic ear detection. In: Seventh International Conference on Advances in Pattern Recognition, 2009. ICAPR'09, pp. 213–216. IEEE (2009).

[5] Bahle, B., & Hollingworth, A. (2019). Contrasting episodic and template-based guidance during search through natural scenes. Journal of Experimental Psychology: Human Perception and Performance, 45(4), 523.

[6] Fu, K., Dou, F. Z., Li, H. C., Diao, W. H., Sun, X., & Xu, G. L. (2018). Aircraft recognition in SAR images based on scattering structure feature and template matching. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(11), 4206-4217.

[7] Fu, K., Dou, F. Z., Li, H. C., Diao, W. H., Sun, X., & Xu, G. L. (2018). Aircraft recognition in SAR images based on scattering structure feature and template matching. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(11), 4206-4217.

[8] Zheng, J., Ranjan, R., Chen, C. H., Chen, J. C., Castillo, C. D., & Chellappa, R. (2018). An Automatic System for Unconstrained Video-Based Face Recognition. arXiv preprint arXiv:1812.04058.

[9] Du, G., Zhou, M., Yin, C., Wu, Z., & Shui, W. (2018). Classifying fragments of terracotta warriors using template-

based partial matching. Multimedia Tools and Applications, 77(15), 19171-19191.
[10] Sinha, H., Manekar, R., Sinha, Y., & Ajmera, P. K. (2019). Convolutional Neural Network-Based Human Identification Using Outer Ear Images. In Soft Computing for Problem Solving (pp. 707-719). Springer, Singapore.
[11] Yang, S., Bo, L., Wang, J., & Shapiro, L. G. (2012). Unsupervised template learning for fine-grained object recognition. In Advances in neural information processing systems (pp. 3122-3130).
[12] Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., & Navab, N. (2012, November). Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In Asian conference on computer vision (pp. 548-562). Springer, Berlin, Heidelberg.
[13] Lee, T., & Soatto, S. (2011). Video-based descriptors for object recognition. Image and Vision Computing, 29(10), 639-652.
[14] Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., & Lepetit, V. (2011, November). Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In 2011 international conference on computer vision (pp. 858-865). IEEE.
[15] Lin, Z., & Davis, L. S. (2010). Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(4), 604-618.
[16] Geng, C., & Jiang, X. (2011). Face recognition based on the multi-scale local image structures. Pattern Recognition, 44(10-11), 2565-2575.
[17] Muyan-Ö zçelik, P., Glavtchev, V., Ota, J. M., & Owens, J. D. (2010, September). A template-based approach for real-time speed-limit-sign recognition on an embedded system using GPU computing. In Joint Pattern Recognition Symposium (pp. 162-171). Springer, Berlin, Heidelberg.
[18] Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., & Lepetit, V. (2012). Gradient response maps for real-time detection of textureless objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(5), 876-888.
[19] Ashtari, A. H., Nordin, M. J., & Fathy, M. (2014). An Iranian license plate recognition system based on color features. IEEE transactions on intelligent transportation systems, 15(4), 1690-1705.
[20] Chae, J., Woo, I., Kim, S., Maciejewski, R., Zhu, F., Delp, E. J., ... & Ebert, D. S. (2011, February). Volume estimation using food specific shape templates in mobile image-based dietary assessment. In Computational Imaging IX (Vol. 7873, p. 78730K). International Society for Optics and Photonics.
[21] Lang, H., Wang, Y., & de Silva Clarence, W. (2010, June). Vision based object identification and tracking for mobile robot visual servo control. In IEEE ICCA 2010 (pp. 92-96). IEEE.
[22] Mahbub, U., Imtiaz, H., Roy, T., Rahman, M. S., & Ahad, M. A. R. (2013). A template matching approach of one-shot-learning gesture recognition. Pattern Recognition Letters, 34(15), 1780-1788.
[23] Anjomshoae, S. T., & Rahim, M. S. M. (2016). Enhancement of template-based method for overlapping rubber tree leaf identification. Computers and Electronics in Agriculture, 122, 176-184.
[24] Bodla, N., Zheng, J., Xu, H., Chen, J. C., Castillo, C., & Chellappa, R. (2017, March). Deep heterogeneous feature fusion for template-based face recognition. In 2017 IEEE winter conference on applications of computer vision (WACV) (pp. 586-595). IEEE.