International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

A Survey Paper on Efficient Object Detection and Matching using Feature Classification

Arjun S Kaushik[1], Soorya Raysam[2], Venkatakrishna S[3], Dr Prabhanjan S[4]

[1,2,3]Dept. of Computer Science, Jyothy Institute of Technology, Bangalore, India
[4]Head of Department, Dept. of Computer Science, Jyothy Institute of Technology, Bangalore, India

---

Abstract - -*This paper presents a new approach for efficient object detection and matching in images and videos. We propose a stage based on a classification scheme that classifies the extracted features in new images into object features and non-object features. This binary classification scheme has turned out to be an efficient tool that can be used for object detection and matching. By means of this classification not only the matching process becomes more robust and faster but also the robust object registration becomes fast. We provide quantitative evaluations showing the advantages of using the classification stage for object matching and registration. Our approach could lend itself nicely to real-time object tracking and detection.*

Key Words: Efficient Object Detection, SSD Mobile-Net, R-CNN

1. INTRODUCTION

The capability of detecting and registering objects in a video sequence captured by either a fixed camera or a moving camera is the corner stone in many computer vision applications. The camera can be a hand-held camera, a robotics camera, or an on board camera. To this end, many challenging problems should be solved, namely object detection, 3D object pose, feature extraction and matching, and image registration. The problem of object detection has been studied by many researchers. Supervised techniques based on learned appearances has been used to detect objects whose class can be described statistically such as faces, facade windows, and vehicle rears. These techniques include Adaptive Boosting and Active Appearance Models. However, in many applications the objects of interest cannot be described by a generic model. For example, tracking an arbitrary physical object cannot use the above techniques. Therefore, the common strategy is to use a reference model for this object. This model can be represented by a template or a set of relevant features. At run time, input images are matched with the object template or features in order to register the object with the current image. The kind of the registration depends on the object in question. Therefore, if the object is planar then the registration aims to compute the homographic transform between a reference frame and the current frame. If the object is 3D then the registration aims to compute its 3D pose or projection matrix with respect to the camera. In all cases, a set of feature matches should be computed before carrying the registration process. The matches can be established using classical feature matching scheme. However, in many cases, one has to overcome a challenge resulting from the fact the object at hand may have a small size in the current captured images.

2. LITERATURE SURVEY

While feature point recognition is a key component of modern approaches to object detection, existing approaches require computationally expensive patch pre-processing to handle perspective distortion. In this paper, it is shown that formulating the problem in a Naive Bayesian classification framework makes such pre-processing unnecessary and produces an algorithm that is simple, efficient, and robust. Furthermore, it scales well to handle large number of classes. To recognize the patches surrounding key points, the classifier uses hundreds of simple binary features and models class posterior probabilities. The problem is made computationally tractable by assuming independence between arbitrary sets of features. Even though this is not strictly true, it is demonstrated that the classifier nevertheless performs remarkably well on image datasets containing very significant perspective changes. [1]

The authors P. Viola and M. Jones describe a visual object detection framework that is capable of processing images extremely rapidly while achieving high detection rates. There are three key contributions. The first is the introduction of a new image representation called the “Integral Image” which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features and yields extremely efficient classifiers. The third contribution is a method for combining classifiers in a “cascade” which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. A set of experiments in the domain of face detection are presented. The system yields face detection performance comparable to the best previous systems. Implemented on a conventional desktop, face detection proceeds at 15 frames per second. [2]

The problem of pose estimation arises in many areas of computer vision, including object recognition, object tracking, site inspection and updating, and autonomous navigation using scene models. A new algorithm, called SoftPOSIT is presented, for determining the pose of a 3D

object from a single 2D image in the case that correspondences between model points and image points are unknown. The algorithm combines Gold's iterative SoftAssign algorithm for computing correspondences and DeMenthon's iterative POSIT algorithm for computing object pose under a full perspective camera model. This algorithm, unlike most previous algorithms for this problem, does not have to hypothesize small sets of matches and then verify the remaining image points. Instead, all possible matches are treated identically throughout the search for an optimal pose. The performance of the algorithm is extensively evaluated in Monte Carlo simulations on synthetic data under a variety of levels of clutter, occlusion, and image noise. The tests conducted shows that the algorithm performs well in a variety of difficult scenarios, and empirical evidence suggests that the algorithm has a run-time complexity that is better than previous methods by a factor equal to the number of image points. The algorithm is being applied to the practical problem of autonomous vehicle navigation in a city through registration of a 3D architectural models of buildings to images obtained from an on-board camera. [3]

In this paper, a local image descriptor is introduced that is inspired by earlier detectors such as SIFT and GLOH but can be computed much more efficiently for dense wide-baseline matching purposes. It retains their robustness to perspective distortion and light changes and can be made to handle occlusions correctly, and runs fast on large images. The descriptor yields better wide-baseline performance than the commonly used correlation windows, which are hard to tune. Too small, they do not bring enough information. Too large, they become vulnerable to perspective variations and occlusion. Therefore, recent methods tend to favour small correlation windows, or even individual pixel differencing and rely on global optimization techniques such as graph-cuts to enforce spatial consistency. They are restricted to very textured or high-resolution images, of which they typically need more than three. This descriptor overcomes these limitations and is robust to rotation, perspective, scale, illumination changes, blur and sampling errors. It produces dense wide baseline reconstruction results that are comparable to the best current techniques using fewer lower-resolution images. [4]

This paper aims to present a review of recent as well as classic image registration methods. Image registration is the process of overlaying images (two or more) of the same scene taken at different times, from different viewpoints, and/or by different sensors. The registration geometrically align two images (the reference and sensed images). The reviewed approaches are classified according to their nature (area-based and feature-based) and according to four basic steps of image registration procedure: feature detection, feature matching, mapping function design, and image transformation and resampling. Main contributions, advantages, and drawbacks of the methods are mentioned in the paper. Problematic issues of image registration and outlook for the future research are discussed too. The major goal of the paper is to provide a comprehensive reference source for the researchers involved in image registration, regardless of particular application areas. [5]

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [7] and Fast R-CNN [5] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, a Region Proposal Network (RPN) is introduced, that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and object scores at each position. RPNs are trained end-to-end to generate high quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model, the detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image.[6]

An auxiliary task to Mask R-CNN is presented along with an instance segmentation network, which leads to faster training of the mask head. The addition to Mask R-CNN is a new prediction head, the Edge Agreement Head, which is inspired by the way human annotators perform instance segmentation. Human annotators copy the contour of an object instance and only indirectly the occupied instance area. Hence, the edges of instance masks are particularly useful as they characterize the instance well. The Edge Agreement Head therefore encourages predicted masks to have similar image gradients to the ground-truth mask using edge detection filters. A detailed survey of loss combinations is conducted and shows improvements on the MS COCO Mask metrics compared to using no additional loss. The approach used marginally increases the model size and adds no additional trainable model variables. While the computational costs are increased slightly, the increment is negligible considering the high computational cost of the Mask R-CNN architecture. As the additional network head is only relevant during training, inference speed remains unchanged compared to Mask RCNN. In a default Mask R-CNN setup, a training speed-up is achieved and a relative overall improvement of 8.1% on the MS COCO metrics compared to the baseline. [7]

Deep Neural Networks exhibit major differences from traditional approaches for classification. They are deep architectures which have the capacity to learn more complex models than shallow ones. This model is capable of predicting the bounding boxes of multiple objects in a given image.To increase localization precision, the DNN mask generation is applied in a multi-scale fashion on the

full image as well as on a small number of large image crops. A multi-scale box inference is presented followed by a refinement step to produce precise detections. In this way, a DNN predicts a low-resolution mask. A single DNN regression can give masks of multiple objects in an image. To deal with multiple touching objects, several masks are generated, each representing either the full object or part of it. Further, if two objects of the same type are placed next to each other, then at least two of the produced five masks would not have the objects merged which would allow to disambiguate them. This would enable the detection of multiple objects. For training the mask generator, several thousand samples from each image divided into 60% negative and 40% positive samples is generated. A sample is considered to be negative if it does not intersect the bounding box of any object of interest. Positive samples are those covering at least 80% of the area of some of the object bounding boxes. [8]

3. METHODOLOGY

Feature point classification or FPC plays an important role in the field of object detection. This is mainly because it eliminates the need of pre-processing the data. But this paper makes FPC a redundant matter. To recognize the patches surrounding key points, the classifier uses many simple binary features and class posterior probabilities. The problem is made computationally tractable by assuming independence between arbitrary sets of features. [1]

A VODF or Visual Object Detection Framework is capable of processing images extremely rapidly while achieving high detection rates. There are three key contributions. The first is the introduction of a new image representation which allows the features used by the detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features and yields extremely efficient classifiers. The third contribution is a method for combining classifiers in a "cascade" which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. A set of experiments in the domain of face detection are presented. [2]

A new algorithm, called SoftPOSIT is introduced, for determining the pose of a 3D object from a single 2D image in the case that correspondences between model points and image points are unknown. The algorithm combines Gold's Soft Assign algorithm for computing correspondences and DeMenthon's iterative POSIT algorithm for computing object pose under a full perspective camera model. All possible matches are treated identically throughout the search for an optimal pose. The performance of the algorithm is extensively evaluated in Monte Carlo simulations on synthetic data under a variety of levels of clutter, occlusion, and image noise. These tests show that the algorithm performs well in a variety of difficult scenarios, and empirical evidence suggests that the algorithm has a run-time complexity that is better than previous methods [3]

SIFT and GLOH, the early detectors inspired the creation of a new local image descriptor which is as robust and efficient as its predecessors.The descriptor yields better wide-baseline performance than the commonly used correlation windows, which are hard to tune. Too small, they do not bring enough information. Too large, they become vulnerable to perspective variations and occlusion. Recent methods tend to favour small correlation windows, or even individual pixel differencing and rely on global optimization techniques such as graph-cuts to enforce spatial consistency. They are restricted to very textured or high-resolution images, of which they typically need more than three. The descriptor used in this paper overcomes these limitations and is robust to rotation and perspective, scale, [4].

The process of overlapping 2 or more images of the exact same scene but taken at different times and different views is called Image registration. The registration geometrically align two images (the reference and sensed images). The reviewed approaches are classified according to their nature and according to four basic steps of image registration procedure: feature detection, feature matching, mapping function design, and image transformation and resampling. One of the main objectives of the paper is to provide a comprehensive reference source for the researchers involved in image registration,.[5]

An RPN or Region Proposal Network is a network that shares complete-image features with the detection network thereby reducing cost on region proposals. An RPNs are trained end-to-end to generate high quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. [6]

R-CNN proved to be slow whilst training the mask head.

So, a new head called the Edge Agreement Head is added to overcome this disadvantage and for instance segmentation. Human annotators copy the contour of an object instance and only indirectly the occupied instance area. The Edge Agreement Head enables predicted masks to have similar image gradients to the ground-truth mask using edge detection filters. The approach used increases the model size and adds no additional trainable model variables. The computational costs are increased slightly. But the increment is negligible considering the high computational cost of the Mask R-CNN architecture. As the additional network head is only relevant during training, inference speed does not change compared to Mask R-CNN. [7]

DNN or Deep Neural Networks are complicated networks which have the ability to learn more complex models than shallow ones. This model is capable of predicting the bounding boxes of multiple objects in a given image. To deal with many objects that are in contact, a large number of masks are generated, each representing either the full object or part of it. Further, if two objects of the same type are placed next to each other, then at least two of the produced five masks would not have the objects merged which would allow to disambiguate them. This would enable the detection of multiple objects. For training the mask generator, many samples from each image divided into negative and positive samples is generated. Depending on whether the image intersects or interferes with the bounding boxes, the sample is classified as a positive sample or negative sample.[8]

4. CONCLUSION

From the above discussions we are able to conclude that though many algorithms and models have been implemented most of them have given high accuracy but slow or low accuracy but faster in the detection of objects. Therefore, in the proposed system we use the SSD model for which the comparison is done to find which model provides the best result.

REFERENCES

[1] M. Ozuysal, P.Fua and V. Lepetit “Fast key porint recognition in Ten Lines Of Code”.

[2] P. Viola and M.Jones “Robust Real-time Object Detection”.

[3] R. Duraiswami and H. Samet “SoftPOSIT: Simultaneous Pose and Correspondence Determination”.

[4] P. Fua and V.Lepetit “A Fast Local Descriptor for Dense Matching”.

[5] B. Zitova and J.Flusser “Image registration Methods:a Survey”.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun “Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Network”.

[7] Ronald and S. Zimmermanna, Julien N. Siemsa “Faster Training of Mask R-CNN by Focusing on Instance Boundaries.

[8] Christian Szegedy , Alexander Toshev,Dumitru Erhan “Deep Neural Network for Object Detection”.

[9] www.google.com