International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206540

# Survey on Real Time Application of Computer Vision

**Utsho Chakraborty[a], Haimanti Biswas[b], Dr. Sheshang Degadwala[c]***
[a]UG Scholar, Computer Engineering, Sarvajanik College of Engineering and Technology, Surat, India
[b]UG Scholar, Information Technology, Sigma Institute of Engineering, Vadodara, India
[c] Associate Professor & Head of Department, Computer Engineering Department, Sigma Institute of Engineering, Vadodara, India

**Article Info**
Volume 6, Issue 5
Page Number: 185-193
Publication Issue :
September-October-2020

**Article History**
Accepted : 20 Sep 2020
Published : 05 Oct 2020

**ABSTRACT**

Computer Vision system (existing) are being used in recent times, measuring what, when, where and how things move in street and open spaces, concerning the fact of promoting public security. Since a multi-dimensional area of learning, it may look messy, with techniques taken from and reused many a times for a range of contrasting engineering and computational science fields; therefore it's the mechanized extraction of information (3D models, camera position, object detection and recognition to grouping and searching image content) originated through images. Our investigation proposes a comprehensive survey on some crucial problems, e.g., Scene Recognition, Action Recognition, Visual Saliency Estimation, Objectness Estimation - focusing on computer vision based perspective.

Keywords: Computer Vision, Multi-dimentional, Mechanized Extraction, Scene Recognition, Action Recognition, Visual Saliency Estimation, Object-ness Estimation.

## I. INTRODUCTION

Computer Vision itself is the amalgamation of ideas, methods, plans and designs through digital image processing, pattern recognition, artificial intelligence and computer graphics, therefore relating to the procedure of gathering information either individually or in a group manner, starting from gathering input scenes (digital images) to trademark extraction. The Image Processing contrarily signifies the mechanization of computational transformations on images in particular, i.e. sharpening, contrast balancing. But as per performance, Computer vision is application dependent. In general, the performance evaluation takes part in calculating some crucial manners of an algorithm in order to achieve accuracy, strength, or extensibility to control and monitor system performance. Moreover, its perspective is used to detect and prevent some specific types of occurring powered by deep learning algorithms focusing on Convolution Neural Networks (CNN) for acquiring, processing, analyzing, and understanding of an image.

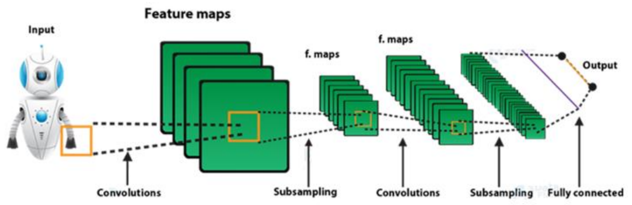



**Fig.1.** Structural Convolution Neural Networks[CNN]

From the prominent figure we find, as a profound neural organization, CNN can catch complex highlights from picture information. In times we find upgrades in CNN techs which is shown below in table.

Table 1 : Constant changes in CNN models through times

| Networks | Density | Technologies |
|---|---|---|
| ZF Net | 5 | width got increased |
| VGG | 16-19 | kernel type reduction is 3x3 size |
| Google Net | 22 | auxiliary classified module |

Moreover the related work focuses on the proposed investigation, which acquired some crucial faults concerning the specifics, leading to some grounds:

- Unprecedented detection of the most common scene type, this visual input belongs to: airport, stadium, indoor, etc.


**Fig. 2.** Types of Indoor Scenes

- Having millions of images and videos on the web, and most of them subjecting *'someone working something'*, resulting in a big need for detecting and recognizing the type of action occurs in those images and videos. Being a trending research agenda, there is still so much work to do.

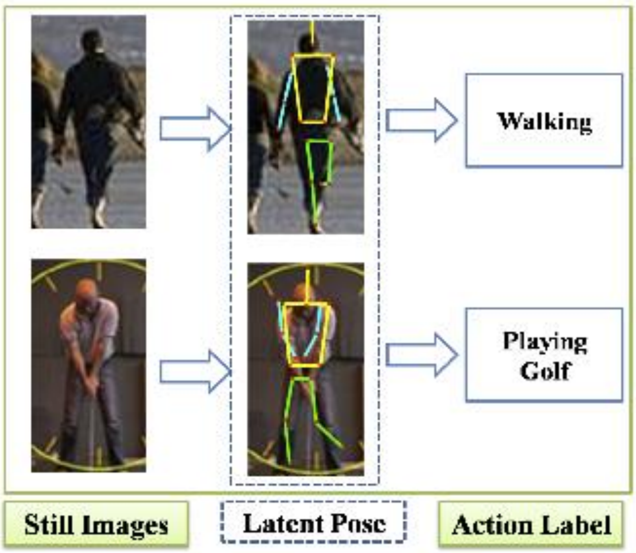



**Fig. 3.** A person's image; aim is to action detection

- Traditionally, detecting an object in an image was to slide a window over all possible areas in all possible scales; which thus resulted in huge computation costs and many false positives: given an image of a cellular size˜2.5-3 million windows to search for an object.

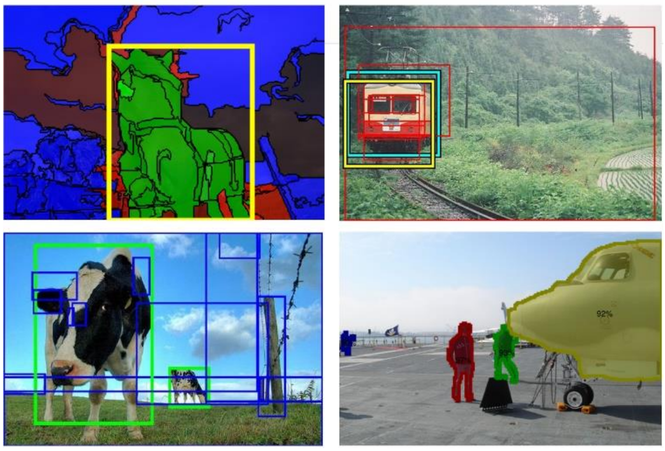


**Fig. 4.** Provided output set of candidate windows and searching objects

SIFT (Scale Invariant Feature Transform) being an algorithm used to detect and describe local features (key points with quantitative information (descriptors)) in digital images. But we find SURF (speedup Robust Feature) is more exclusive. It can be used under tasks such as object recognition, image registration, classification, or 3D reconstruction.

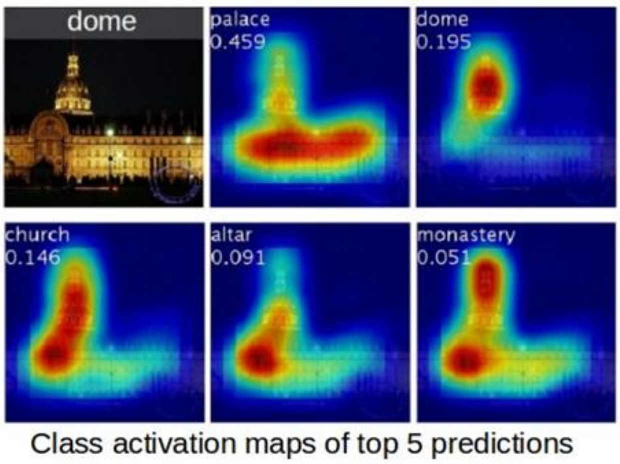
**Real Time Application of Computer Vision :**

Table 2: Difference within the Traditional and Proposed Approach

| Perameter | Traditional | Proposed |
|---|---|---|
| Transform | Simple Vision | Complex Vision |
| Feature approach | Local (SIFT/SURF approach approach) | Unprecidented |
| Action Recognizer | Local Cache | Specifier Unrecognized |
| System | Automatic | Upgraded |
| Covering Image | Grey | Multicolored (green, red, blue, etc.) |

The transformation here from the primary approach, we find that Computer Vision following with CNN versions goes through the Simple vision approach for detecting purposes, whereas the proposed approach finds it complexion and therefore yet to be solved, as the features in the traditional approach were local findings(SIFT/SURF); therefore local image processing was a trend, which in recent times became much crucial with identification of every individual type, making it quite unprecedented. The System was as automated as it has been in the recent decade but due to constant upgrades in tech's it has become much more crucial for security, better system performance.

**Class activation & specific Saliency Mapp Prompts**



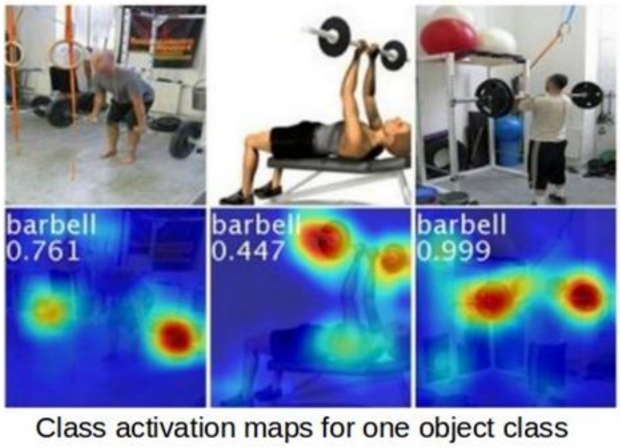




**Fig. 5.** Class Activation & specific Salience Mapping

Class enactment guides could be utilized to discuss the expectation choice made by the CNN. The left picture underneath shows the class enactment guide of top 5 forecasts separately, you can see that the CNN is set off by various semantic areas of the picture for various expectations. The correct picture underneath shows the CNN figures out how to confine the regular visual examples for a similar article class. Moreover, the profound highlights from our organizations could be utilized for conventional limitation, with recently prepared SVM's loads to create the class actuation map, at that point you could get class-explicit saliency map for nothing.

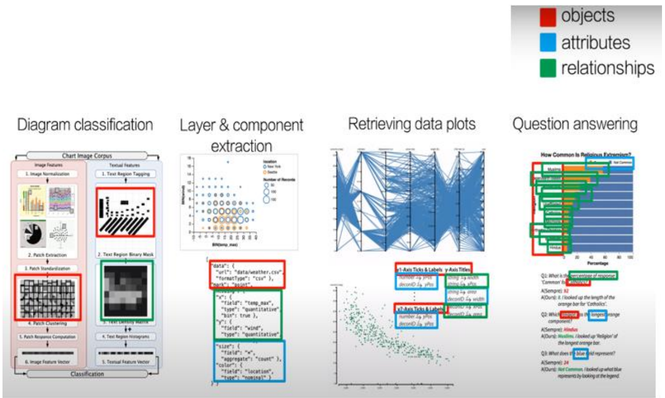
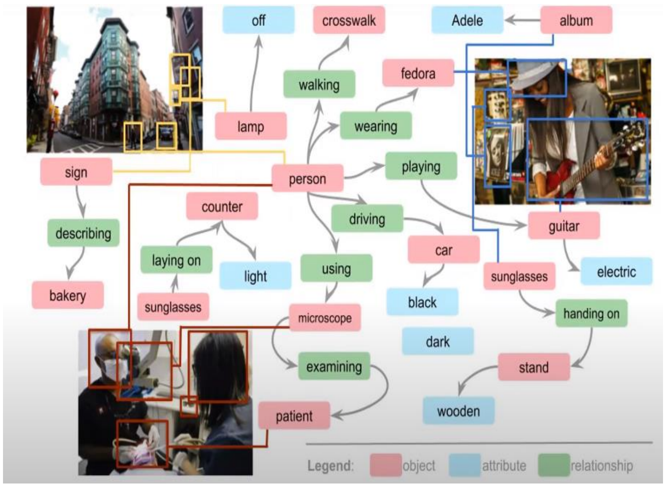



**Fig.6.** Symbolic representation of structures

The Structure represents the Tasks in which people work on diagrams, data plots, and charts for decades, expanding which, has been eventually quite useful for the users, where the diagrams were not only classified within various categories but also building models for extraction of the components from them, retrieving data plots by user search and modification of the underlying structure of those plots, even answering open-ended questions concerning diagrams.

Throughout time, when these tasks became more useful; they also became more complex, requiring more datasets, human interactions, and complicated modeling choices. Significantly, a theme from all these tasks here is this underlying structure, where all the tasks models must learn in recognizing grounds, elements, objects for the data entities i.e. Patches, Labels, etc. Additionally, they need to acknowledge the attributes associated with each entity i.e. Size, Color of a point, Numbers, etc. and finally the relationship between those entities i.e. the absolute and relative spatial locations between the entities and their associations between categories and values, the density between patches.




108k Images
3.8Million Attributes
2.8Million Attributes
2.3 Million Relationships

**Fig.7.** Visual genome – Connecting the images!

The figure here, however, finds the values of Visual Genomes. **Visual Genome,** a dataset, a knowledge base, is an ongoing effort to connect structured image concepts to language i.e. playing Guitar, helping, Marking people, and other objects. The very first visual Genome Dataset was released in 2017 with several 108k Grapes, with millions of objects, attributes, and relationships, allowing them to connect all and having a systematic workflow and this result has been more fruitful ever before.

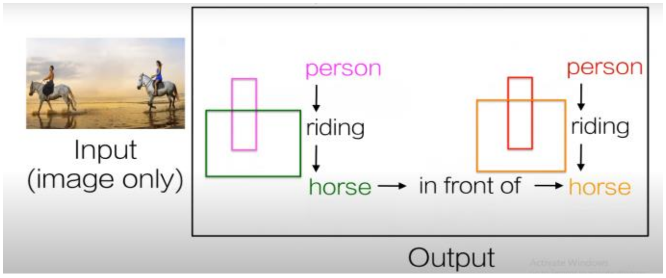



**Fig.8.** Scene graph Generation

To promote the sharing of interdisciplinary ideas i.e. intricate underlying structure or networks that define the interaction with the components; we profound the scene graph generation where, given any given input, models must produce a structured representation of those image inputs and thus encoding the relationships between all these grounded objects, formulating the necessary outputs. The given graph symbolizes the same fact where we find the person riding a horse where a relation is in front of!

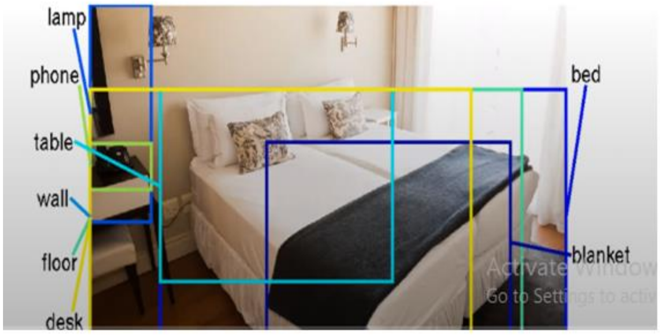



**Fig.9.** Improved results of Image Retrieval

The figure represents the improved results through Image Retrieval utilizing scene graphs. Instead of the retrieval of the component, we achieved the platform of object retrieval, throughout the generations over the last couple of years, were given any long complex sentenced description, resulting in a stable conversion of scene graph and grounding it in training/retrieval set and retrieve the right images for the given query of sets.

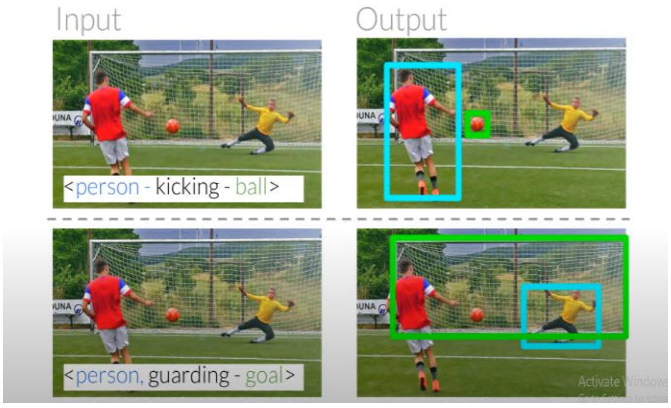



**Fig.10.** Specific Object localization (vision task)

Understanding/ Modeling the relationships while having proper retrieval of the specific objects given, can also improve object localization. Given that the detection of a specific object if very small i.e. ball from the image, is very hard to do, knowing that a person is kicking the ball, provides enough contextual clues that folk models ability to reason about where the ball must be. Similarly, these techniques help us enabled instance localization, where the model localizes people differently depending on whether the person is kicking the ball for or guarding the Goal.

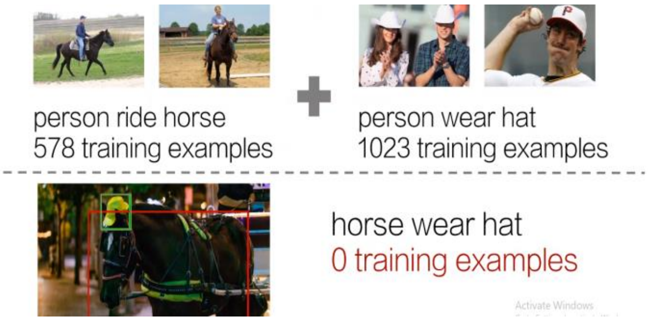



**Fig.11.** Compositionality through Object collaboration (predictability)

One of the most important aspects is the compositional structure that the graph provides, enabling few-shot prediction by looking at examples of people riding horses while wearing hats, through which we can now build models composing this individual objects and relationships for detection of completely novel compositions i.e. being able to detect when a horse is wearing a hat even though it might have never seen that in its training set.

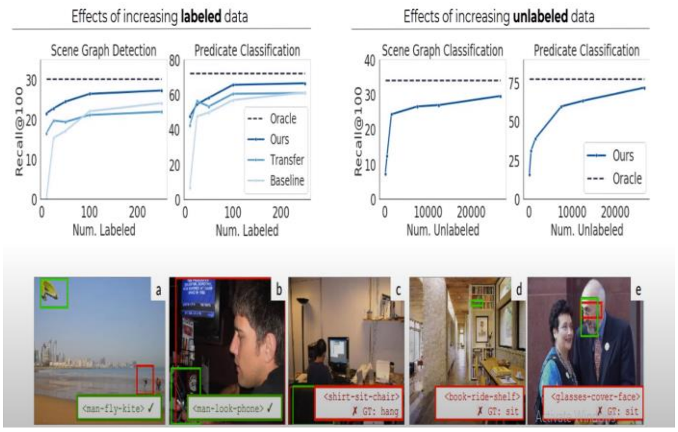



**Fig.12.** Performance Improvement with Diminishing Gains based on increase instances (un/labeled)

Oracle measuring training on entire visual genome dataset, we get to find through our mistake approach that, when the categorical or spatial features were insufficient to explain the variability of a relationship i.e. in Part C here, we find the model predicts that the shirt is sitting on a chair when it should have predicted hanging on a chair but it has only seen people hanging on bars whereas it should be people are hanging below the bars or so on. therefore, it is so hard to generalize to a situation where the shirt is hanging on something whilst performing that action it is on top of the thing it's hanging from.

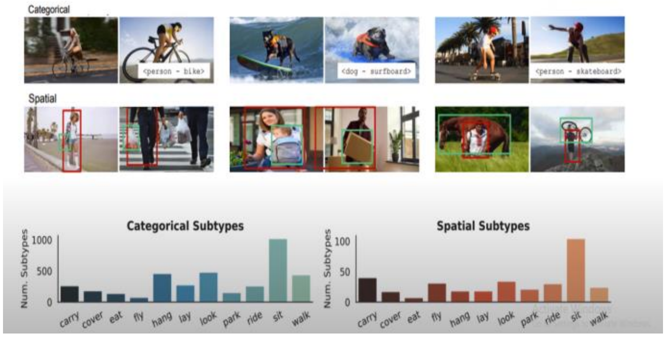



**Fig.13.** Relationship Complexity within subtypes

The importance of representing this variability of a relationship we defined a notion of complexity where a relationship participates in more relationships with more categories in more spatial orientations which shows the significant change throughout the relationships and we find

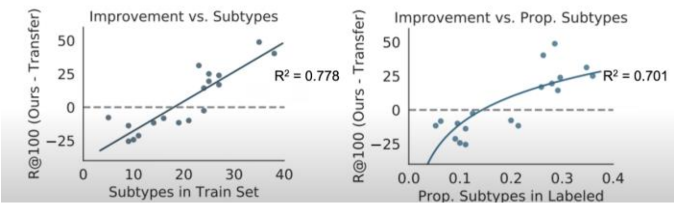



**Fig. 14.** Subtypes in the small labeled set concerning improvement

that the models performance is in fact directly co-related with no's of subtype or complexity of that relationship, suggesting that some relationships are quite easier to model than others and that if we were to build a few more learning methods for relationships, we need to understand the complexity of that relationships and making sure that we have a representative enough size in our labeled set that can thus generalize to a lot of unlabeled images.

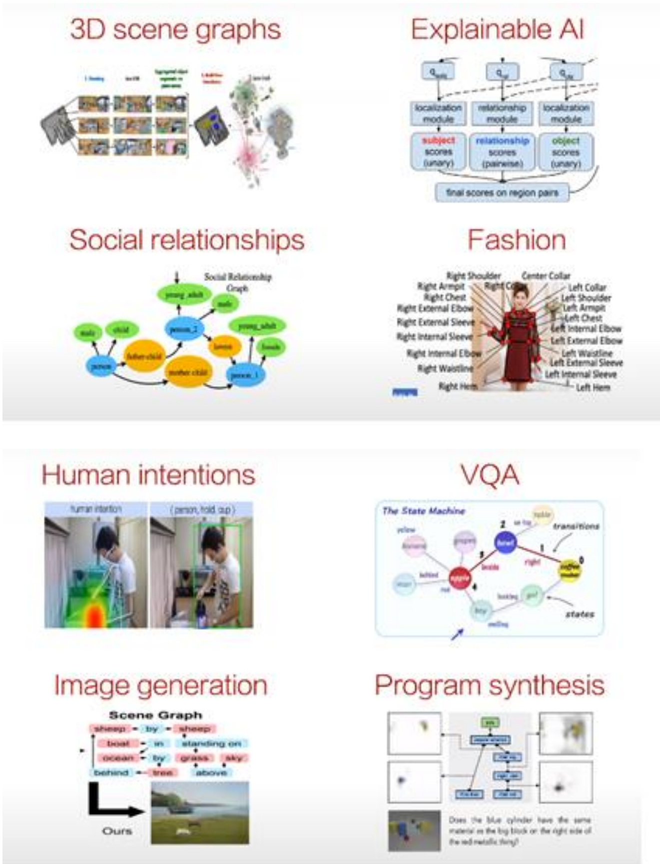



**Fig. 15.** A state of Art in diverse fields- Graphical Essences!

Now a days the task of graph's generations has inspired models to go beyond just generating scene graphs but utilizing these generated scene graphs for a whole host of core computer vision tasks i.e. 3D understanding, Explainable AI, Understanding Human Intentions, Social Relationships, Generating images and programs and answering questions

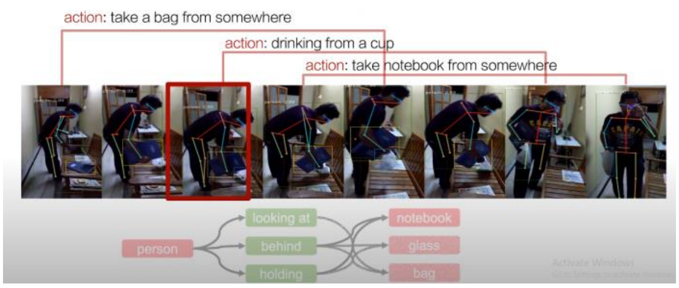



**Fig.16.** Action Genome: Understanding Action in Spatio-temporal prospect

Extending the Scene Graph within the temporal Domain as Spatiotemporal Scene Graphs, as they tend to change over time as people perform actions. For example, while moving on further on the scene, the graph itself evolves over time.

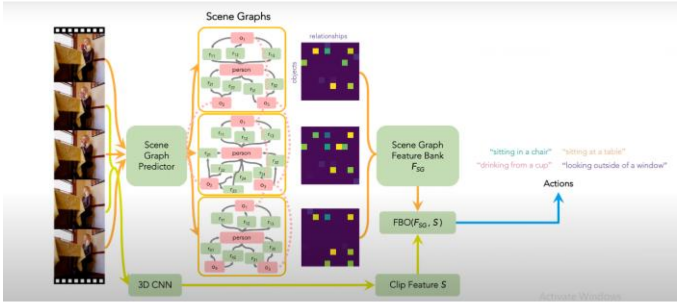



**Fig.17.** Spatio-temporal Scene Graph feature Banks (SGFB) focusing Action Recognition

A method for a few shots of action recognition using existing scene graph generation models for generating these scene graphs for every frame of the video. Instantly, for every graph, the model starts by generating these scene graphs for individual frames and thus by encoding all graphs within latent representation, thus fuses them within a feature bank along another branch, were we also encode 3D CNN features and thus combining and fusing with feature bank to finally predict the actions that are taking place at different points.

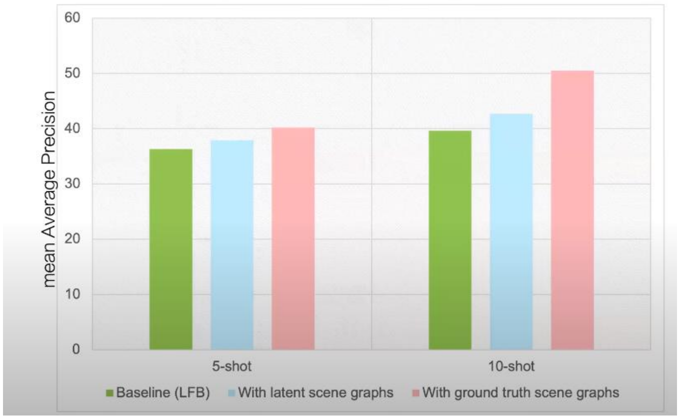



**Fig.18.** Spatio-temporal Scene Graph improve few-shots action classification

Explicitly modeling the scene Graphs can lead to improvements in detecting actions against recently proposed strong baselines using as few as 5 to 10 examples per actions, expecting the growth of no's on future works incorporating these Spatio-temporal transformations of these graphs into better action recognition models.

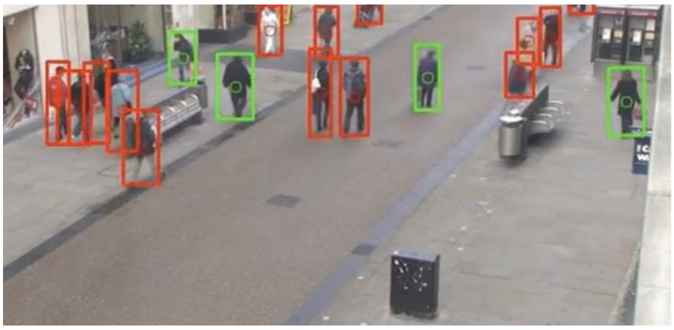
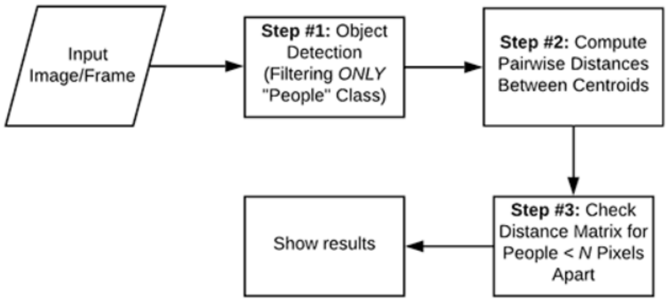


**Fig.19.** Social Distancing Detector concerning COVID19

The following figure opens up a very crucial path quite significant in controlling the spreading of the virus and through object detection, with the help of YOLO Object detector and python driver script, glues all the pieces together in building Open CV social distancing detector, through which computer vision has achieved yet another evolution. There are some certain steps to be followed which are followed below,




## II. CONCLUSION

Our proposed Investigation infers that zeroing in on object discovery, even though Deep Learning has drastically improved all the curves up, there still has been and still is a huge gap between a human and a computer. At the beginning frontiers, humans are very successful in learning diverse things, that too only with couple of instances! which actually may be contrary to state of art recognition models that are relying mostly on "Big Data". Additionally, human can extrapolate by creating abstractions in mind, i.e. one seeing something from a particular view point, can still recognize that even if the object proceeds a little or changes its position and standing, therefore again this becomes one of the imperfection of Object Recognition models. As shown recently, even changing a pixel value can misguide state of art recognition model very drastically.

Un/Semi-supervised study of visual things are also a trending factors to dealt with; where Humans are able to learn things with/out labels. Moreover, doing it simultaneously, i.e. knowing that dogs are four legged, and when we see a cat, funny though but without knowing it, we can guess at least, it is an animal. We can also add that human beings learn things by detecting patterns with/out relations and are able to label them. Contrarily, the continuity of the learning remains, i.e. a little child, doesn't say let's learn stuffs. Rather, it's brain progressively processes information, detects design patterns, resulting in findings of many problems needed to be unfolded and materialistically, studying these are quite fun but very hard to solve. What we believe is, Deep Learning or other counterparts are not the major problems here. In fact, they are at present quite successful solutions but we should focus on more to exact problems like the ones we stated above.

## III. REFERENCES

[1]. CONG T. NGUYEN1,5,6, YURIS MULYA SAPUTRA1,3, NGUYEN VAN HUYNH1 , NGOC-TAN NGUYEN1,7, TRAN VIET KHOA1,4, BUI MINH TUAN1,4, DIEP N. NGUYEN1 , DINH THAI HOANG1 , THANG X. VU2 , ERYK DUTKIEWICZ1 , SYMEON CHATZINOTAS2 , AND BJÖRN OTTERSTEN.2 , "A Comprehensive Survey of Enabling and Emerging Technologies for Social Distancing — Part I: Fundamentals and Enabling Technologies"

[2]. C. T. Nguyen, Y. M. Saputra, N. V. Huynh, N. T. Nguyen, T. V. Khoa, B. M. Tuan, D. N. Nguyen, D. T. Hoang, T. X. Vu, E. Dutkiewicz, S. Chatzinotas and B. Ottersten, "A comprehensive survey of enabling and emerging technologies for social distancing — Part II: Emerging technologies and open issues," IEEE Access, 2020

[3]. World Health Organization, "Timeline of WHO's response to COVID19." World Health Organization, June 29, 2020. Accessed: June 29, 2020. Online]. Available: https://www.who.int/news-room/detail/29-06-2020- covidtimeline.

[4]. World Health Organization, "Coronavirus (COVID-19)." World Health Organization, Apr. 22, 2020. Accessed: Apr. 20, 2020. Online]. Available: https://covid19.who.int/

[5]. Mao Jiafa, Wang Weifeng, Hu Yahong & Sheng Weiguo , "A scene recognition algorithm based on deep residual network" 7:1, 243-251, DOI: 10.1080/21642583.2019.1647576 Available : https://doi.org/10.1080/21642583.2019.1647576

[6]. Andrews, S., Tsochantaridsi, I., & Hofmann, T. (2002). Support vector machines for multiple-instance learning. In Advances in neural information processing systems (pp. 561–568)

[7]. Ding, X., Luo, Y., Li, Q., Cheng, Y., Cai, G., Munnoch, R., ... Wang, B. (2018). Prior knowledge-based deep learning method for indoor object recognition and application. Systems Science & Control Engineering, 6(1), 249–257.

[8]. The present and future scope of Computer Vision ~Forbes.com , available blog: https://www.forbes.com/sites/cognitiveworld/2019/06/26/the-present-and-future-of-computer-vision/#7035ee47517d

[9]. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba., “Learning Deep Features for Discriminative Localization”. CVPR'16 (arXiv:1512.04150, 2015).

[10]. JASON BROWNLEE, “A Gentle Introduction to Computer Vision”. (March19, 2019) in DEEP LEARNING FOR COMPUTER VISION˜ Machine Learning Mastery, available Blog website: https://machinelearningmastery.com/what-is-computer-vision/#:˜:text=Computer%20vision%20is%20a%20field,infer%20something%20about%20the%20world.&text=The%20goal%20of%20computer%20vision%20is,the%20content%20of%20digital%20images.

[11]. Samajpati, Bhavini J., and Sheshang D. Degadwala. "A survey on apple fruit diseases detection and classification." International Journal of Computer Applications 130.13 (2015): 25-32.

[12]. Parmar, Pooja A., and Sheshang D. Degadwala. "A feature level fusion fingerprint indexing approach based on MV and MCC using SVM classifier." 2016 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2016.

[13]. Degadwala, Sheshang D., and Sanjay Gaur. "A study of privacy preserving system based on progressive VCS and RST attacks." 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). IEEE, 2016.

[14]. Parmar, Pooja A., and Sheshang D. Degadwala. "Fingerprint indexing approaches for biometric database: a review." International Journal of Computer Applications 130.13 (2015).

[15]. Rana, Ms Dipixa H., and Mr Sheshang D. Degadwala. "Various Techniques of Image Fusion." IJSRD-International Journal for Scientific Research & Development| 2.10 (2014): 2321-0613.

[16]. Samajpati, Bhavini J., and Sheshang D. Degadwala. "Hybrid approach for apple fruit diseases detection and classification using random forest classifier." 2016 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2016.

[17]. Degadwala, Sheshang D., and Sanjay Gaur. "4-Share VCS Based Image Watermarking for Dual RST Attacks." Computational Vision and Bio Inspired Computing. Springer, Cham, 2018. 902-912.

[18]. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440. 2015. DOI: 10.1109/CVPR.2015.7298965

**Cite this article as :**

Utsho Chakraborty, Haimanti Biswas, Dr. Sheshang Degadwala, "Survey on Real Time Application of Computer Vision", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 185-193, September-October 2020. Available at doi : https://doi.org/10.32628/CSEIT206540
Journal URL : http://ijsrcseit.com/CSEIT206540