SPATIAL COMPUTING: CONNECTING PERIPHERAL ACTIVITY THROUGH OBJECT DETECTION USING WEBCAM

[1]AJAY DABAS, [2]ANKIT SHUKLA, [3]PRATYUSH AVI, [4]MANISH PALIWAL, [5]JITENDRA MADARKAR

[1,2,3,4,5]Dept. of Information Technology, Narsee Monjee Institute of Management Studies, India
E-mail: [1]ajay.dabas012@nmims.edu.in, [2]ankit.shukla060@nmims.edu.in, [3]pratyush.avi004@nmims.edu.in, [4]paliwalmanish1@gmail.com, [5]jitendramadarkar475@gmail.com

Abstract - Deep Learning techniques are able to handle several tasks and as well as complex data from various sources better than previous state-of-the-art techniques and capture intricate structures of large scale data by understanding multi-modal information. Therefore, it is important in today’s world, and it has diverse applications such as facial recognition, medical images, object tracking, and so on. Consequently, for computer vision, deep learning techniques have shown to perform the most effectively. This paper highlights the early models and technologies used for object detection and image processing as well as the current advancements in the domain using opensource technologies. It aims to facilitate further research and development into creating hybrid environments that combine real and virtual environments to realise Spatial Computing applications for connected peripheral devices like the keyboard, mouse and trackpad for general purpose usage in various fields, as well as the viability of doing so as a result of technological advancements that have led us to the present technologies.

Keywords - Deep Learning, Computer Vision, RCNN, Virtual Keyboard, Graphical Keyboard.

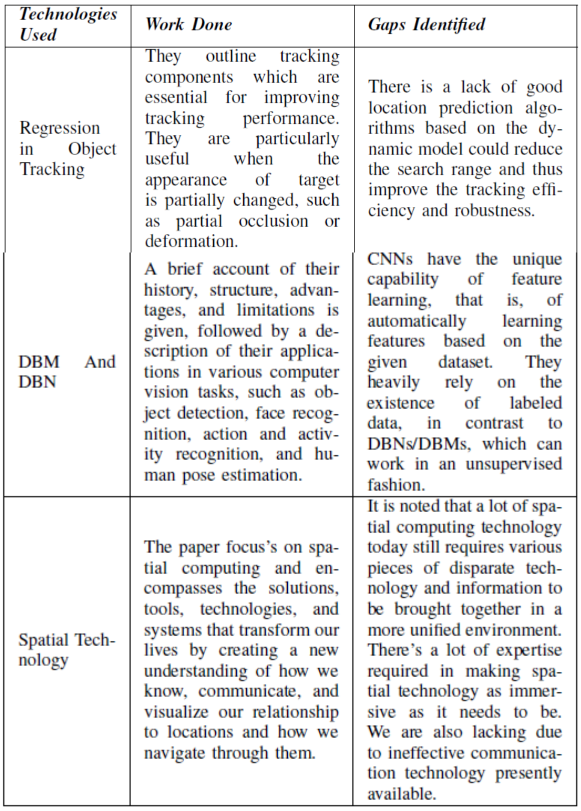
I. INTRODUCTION

Deep Learning techniques have been the most effective when employed for computer vision, which has proved to be the most rapidly evolving and employed subset of it [1]. It has allowed scientists to develop new applications which has been an evolutionary success such as facial recognition [15] [16] [17] and machine learning. This paper aims to facilitate research and development into the next generation of computing which is Spatial computing [2], combining real and virtual environments, starting from basics, and moving to more advanced models. It demonstrates how making such applications are now viable with present technologies and that the concept of Spatial Computing is not an arcane one anymore. We develop and implement an object drag and drop model as well as a gesture-based zoom control application. In addition to this, we also build a virtual keyboard system that can be commonly used by developers and scientists to reduce usage of physical keyboard and ease allay the restrictions posed by hardware for defining computing environments. This paper presents the technologies and methodology for the developed applications which employs finger joints tracking, hand tracking, gesture recognition and artificial marker tracking.

A video sequence is used to capture the movements of the user’s hands. This study shows that it is completely viable to detect and recognize up to 28 keys using real-time video data. The results motivate the development of vision-based virtual keyboards for mobile devices and general inclination to the field of Spatial Computing. Everything brought about by computers has changed and revolutionized many aspects of life around us. With our project we aim to revolutionize computing environments. It is a norm to include physical components such as keyboards, mice, trackpads to establish a functional working environment. The paper seeks to provide perspective to an alternate way of establishing computing environments which may eliminate the need of physical components entirely in near future. With the present open source technologies, it is viable to manage substitutes for components like the keyboard and mice. In case a component fails, one can utilize methods employed in this paper using OpenCV and RCNN to act as substitutes instead of jeopardizing the task that was being carried out. The outlined methods will enable the amalgamation of the virtual and real environments and help carry out tasks like their regular counterparts.

II. LITERATURE SURVEY

The objective of the implemented models i.e., the virtual keyboard, virtual drag and drop and the gesture-based zoom control is to describe the analysis, accuracy, feasibility, of making next generation spatial computing environments that can possibly eliminate the need for physical computing environments altogether. After reviewing the few papers available in the field, shortcomings of previously implemented techniques and models are noted and the viability and practicality of implementing more efficient models using open-source technology is explored. Earlier models on object tracking are noted to be heavily derived from regression techniques and its derivatives. We implement CNN and more specifically RCNN which helps us attain acceptable performance and throughput. [8] helps us reference the performance of implemented models and we realize that the throughput of models implemented on current technology is not sufficient to implement independent alternatives of peripheral devices like the mouse and trackpad but rather make substitutes of such devices. Early prototypes and applications to implement the combination of virtual and real environments have been explored in the past. The most prominent model which could not garner mainstream traction due to inadequate crowd funding and wanting interest from the scientific community is the Leap Motion controller. Other alternatives also include Google Scale, Cardboard, Forge, 360°media. While they do not provide stand-alone devices like Leap Motion, they provide software development kits. These kits are digitally available and thus, more sustainable for businesses than physical devices offered by Leap Motion. The controller was intended as a speciality device with an integrated camera and software which enabled dynamic hand gesture recognition [10]. It most closely resembles the ideology that this paper intends to follow i.e., facilitate the direction of research and development into creating hybrid environments that combine real and artificial environments. The device's primarily functionality is now readily available via the open-source OpenCV software which substantiates the vast technological advancement that warrants a rekindled interest in the field. It was possible to overcome this limitation solely due to the huge development community that updates, refines and contributes to the betterment of OpenCV.



Technologies Used	*Work Done*	*Gaps Identified*
Regression in Object Tracking	They outline tracking components which are essential for improving tracking performance. They are particularly useful when the appearance of target is partially changed, such as partial occlusion or deformation.	There is a lack of good location prediction algorithms based on the dynamic model could reduce the search range and thus improve the tracking efficiency and robustness.
DBM And DBN	A brief account of their history, structure, advantages, and limitations is given, followed by a description of their applications in various computer vision tasks, such as object detection, face recognition, action and activity recognition, and human pose estimation.	CNNs have the unique capability of feature learning, that is, of automatically learning features based on the given dataset. They heavily rely on the existence of labeled data, in contrast to DBNs/DBMs, which can work in an unsupervised fashion.
Spatial Technology	The paper focus's on spatial computing and encompasses the solutions, tools, technologies, and systems that transform our lives by creating a new understanding of how we know, communicate, and visualize our relationship to locations and how we navigate through them.	It is noted that a lot of spatial computing technology today still requires various pieces of disparate technology and information to be brought together in a more unified environment. There's a lot of expertise required in making spatial technology as immersive as it needs to be. We are also lacking due to ineffective communication technology presently available.

III. PROPOSED METHODOLOGY

The models implemented in the study utilize the concept of CNN and specifically RCNN which is region-based CNN. Faster region based convolutional neural network algorithm is used to propose the hand region and the objects that are to be considered in the region i.e., the fingers and palm in this case. The hand tracking module which is the common module utilized by all models in the study, localizes the hand and palm from the input stream. The models then differentiate in their functions. The virtual keyboard model registers inputs based on the distance between the tips of the specified fingers on either hand and prints the output on the output stream.

The gesture-based zoom control model registers the specific gesture of extending only the index finger and the thumb of both hands from the input stream to then enable object size manipulation in the defined virtual space on the output stream i.e., to enlarge or minimize the object on screen. The virtual drag and drop model too recognizes the extended index and middle finger of each hand to activate object selection, dragging the object around the defined space virtually and dropping it at any desired location in the virtual space.

A. Object Detection

Object detection is the ability of computer and software systems to locate and identify items in an image or scene. The act of defining items within photographs normally entails printing bounding boxes and labels for each object. This varies from the classification/localization task in that classification and localization are applied to a large number of objects rather than a single dominant object.

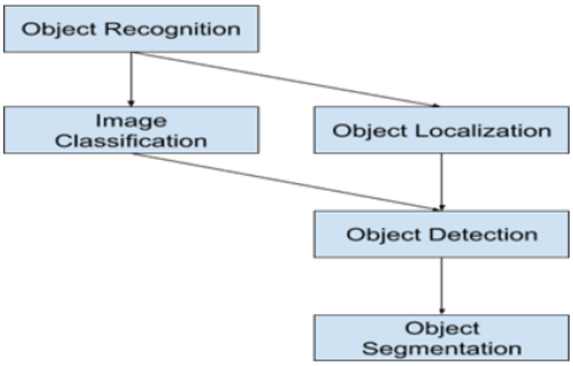


Fig. 1. Schema of Object Detection.

B. R-CNN

The R-CNN method is part of the R-CNN family of algorithms, which stands for ”Regions with CNN Features” or ”Region-Based Convolutional Neural Network.” It's possible that it was one of the first large-scale and successful applications of convolutional neural networks to the problem of object detection, segmentation, and localization. On the VOC-2012 dataset and the 200-class ILSVRC-

2013 object detection dataset, the technique was shown on benchmark datasets, yielding state-of-the-art results [5]. Their proposed R-CNN model is comprised of three modules; they are:

- Region Proposal - Produce category-independent region proposals, such as candidate bounding boxes, then extract them.
- Feature Extractor - Using a deep convolutional neural network, extract features from each candidate region.
- Classifier - Use a linear SVM classifier model to classify features into one of the known classes.

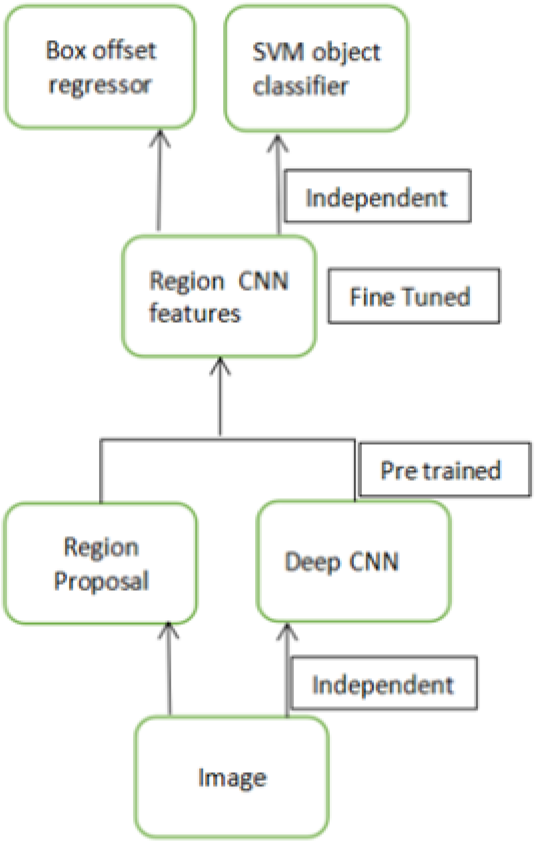


Fig. 2. R-CNN Workflow

C. Fast RCNN

Instead of a pipeline, Fast R-CNN is offered as a single model that learns and outputs regions and classifications directly. The model's architecture takes a photograph as input and generates a set of area recommendations, which are then processed by a deep convolutional neural network. For feature extraction, a pre-trained CNN, such as a VGG-16, is employed. A unique layer termed a Region of Interest Pooling Layer, or Roi Pooling, is added at the conclusion of the deep CNN to extract features relevant to a given input candidate region. The CNN's output is then processed by a fully connected layer, and the model splits into two outputs: one for class prediction via a SoftMax layer, and another for bounding box prediction via a linear output. This technique is then done for each region of interest in a given image many times.

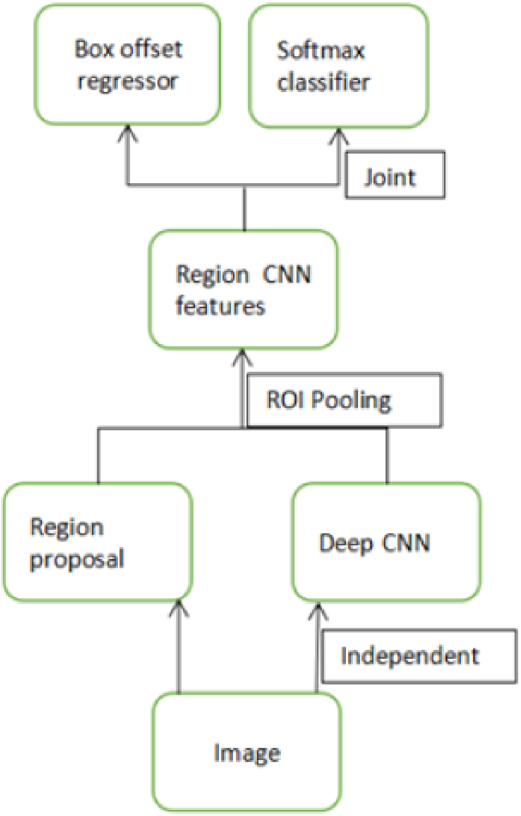


Fig. 3. Fast R-CNN Workflow.

D. Faster RCNN

The architecture, known as a Region Proposal Network, or RPN, was created to both propose and refine region proposals as part of the training process. In a single model architecture, these areas are combined with a Fast R-CNN model. These enhancements lower the number of region proposals while also speeding up the model's test-time operation to near realtime, resulting in state-of-the-art performance. Although it is a single unified model, the architecture is comprised of two modules:

- Region Proposal Network - Convolutional neural networks are used to propose regions and the kind of objects that should be considered in those regions.
- Fast R-CNN - Convolutional neural network for extracting characteristics from suggested regions and generating bounding boxes and class labels

Both modules work with the same deep CNN output. The Fast R-CNN network uses the region proposal network as an attention mechanism directing the second network where to look or pay attention.

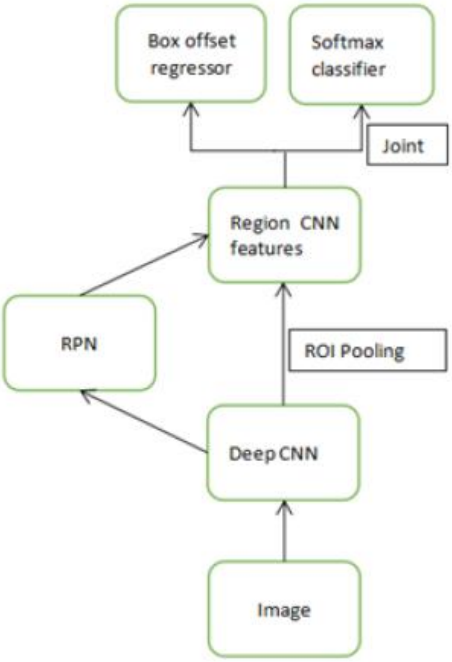


Fig. 4. Faster R-CNN Workflow.

E. Object Tracking

Object tracking is a simple yet challenging task. It is a complex process that involves learning how to identify objects using their pixel values as it does not understand what an image is and what its pixel values are. Machine learning and deep learning techniques can be used to perform object tracking tasks [6]. The former provides us better results on complex problems, while the latter requires a lot of data to complete. This technique finds various applications such as security, surveillance, and traffic monitoring. It can also be utilized for monitoring robots and people. Because of real life applications, the research being conducted in different fields of object tracking focuses on achieving higher accuracy and making the models more robust [7]. The task seems simple for an average human being but it’s way too complex for even the smartest machine. It is not necessary that you create a virtual reality machine, but it is worth keeping an eye on how you do it and how you implement it. The goal here is to provide a simple platform for the user to create virtual reality environments from scratch and it means to get people around the problem and to generate new content for using VR. We employ object tracking in our application by tracking human markers namely the hand, fingers and defined artificial markers which are the objects.

IV. IMPLEMENTATION

Virtual Keyboard project is made at the intersection of computer vision and human-machine interaction. The webcam has a resolution of only 1280 x 720 pixels and thus, becomes a bottleneck.

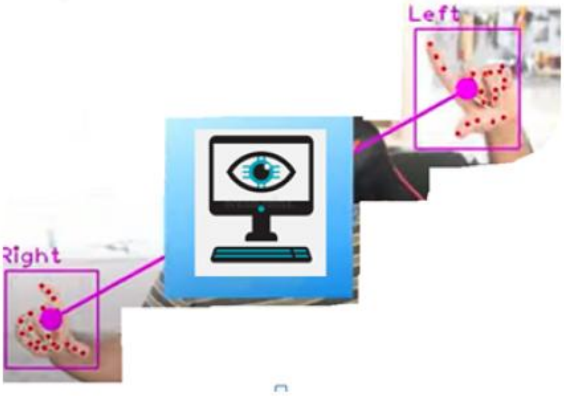


Fig. 5. Zoom In Zoom Out.

The keyboard works seamlessly with computer vision software. The webcam is connected to a remote desktop that aids in monitoring the movements. Using the prototype, it is clearly possible to construct more, and better computer monitors and virtual terminals. The advantage of the virtual keyboard is that it is based on the open-source OpenCV library, and its simplicity and wireless interface makes it easy for users as they do not have to press physical keyboard keys and only show hand gestures to type on keyboard. The software that makes it possible for computer monitors to display and control a mouse, keyboard, or even webcam is OpenCV. OpenCV is one of the most prominent computer vision libraries. It is a programming library that was created to enable programmers to enter the world of computer vision and their library that enables computer vision applications to process images. It is an open-source library that has thousands of functions which are used to shape images. The study has shown various types of filters that play a crucial role in image processing while working on our computer vision applications. OpenCV uses its own inbuilt functions by just writing few lines of code [4]. We will be using OpenCV for applying method works by transforming the raw image into a data/object type. The OpenCV library has more than 2,500 optimized algorithms, which can be used to identify faces, classify objects, produce 3D point clouds, extract 3D models of objects, and detect and remove red eyes from photos. With many established companies such as Google, Yahoo, IBM, and Microsoft, OpenCV is being widely employed by startups.

Some of these include VideoSurf, Applied Minds, and Zeitera. OpenCV is being used to aid in a wide range of applications. It leans towards high performing processors since it supports parallel processing which favours hyper-threading via multi-core processors, taking advantage of SSE and MMX (processor technologies that enable single instruction multiple data). This in turn indicates the readiness of the library to be employed in applications that facilitate spatial computing using current generation processors and network technologies to combine real and virtual environments.

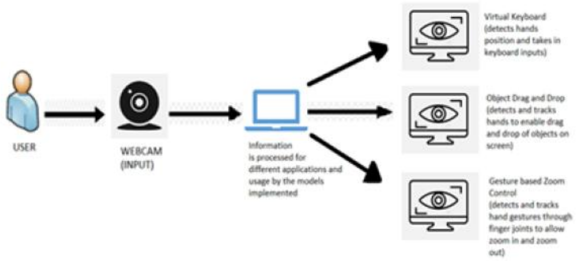


Fig. 6. Methodology for Models Implemented.

V. RESULT ANALYSIS

Post assessing the models, it is clear that there are limitations posed by current technology and, are not ready to phase the hardware components completely as of now due to the increase in throughput of input and output of components such as the keyboard, mice and trackpads. Where the traditional components have a throughput as low as 1ms for very high quality hardware, the models implemented have an average throughput of 1.25s to complete each individual task. This further proves that these models can act as substitutes currently with the scope of improving multifold in the near future through invested research and development in the sector. The currently available hardware poses limitations on the speed and accuracy of completion of tasks, enough to make it non-viable as a primary method of input at this point in time. Despite the potential of the models implemented, these models are very resource intensive to run on sub par hardware which conquer the market of computers at present.

The lack of dedicated research and development in this direction of computer vision shows immense untapped potential of the field. As we overcome the limitation of processing power by the continued progress and innovation, making processing power cheaper every iteration, the possibility of mass-produced speciality devices and software are undoubtedly, endless. OpenCV being an open-source software receives constant updates and support from the open-source developer community and is being revised and refined constantly. It has previously found application in commercial products and forms the backbone of many computer vision products that we anticipate and already see around us. Furthermore, these applications can facilitate new directions of research and development in the field of Computer Vision, re-define how we interact with computing environments and give rise to a very high-potential market for speciality devices and software. Applications can be expected to find wide application in design, academics in the form of interactive learning, diagnostics in healthcare, and employee training in various industry sectors.

VI. FUTURE SCOPE & CONCLUSION

The models employed in the study justify the possibility for major change in the throughput of the models to reduce latency and lag with more powerful and efficient SoCs' in the future. Also, the market for cheaper yet higher quality webcams would go a long way to help such hybrid models which seek to combine the virtual and real environments. The input/output time of models in the future can be reduced to the point their match or surpass their physical hardware rivals (closer to 1ms the better) and can be effectively prioritized as full fledged components of the computer rather than mere temporary substitutes. Much faster and efficient models can be implemented by employing more advanced algorithms than the fast RCNN and faster RCNN that will be widely available in the future. This can be evidently concluded that most of the efforts in the field of computer vision are dedicated to the field of autonomous driving and its sub-systems, and the research and development in the field of spatial computing environments [2] is wanting at best. It is also safe to assume that tangible and spatial awareness interaction systems and applications though still nascent, are progressing rapidly.

OpenCV favours hyper threading since it supports parallelprocessing and therefore the increasing number of cores in every iteration of processors is proving to be favourable to help eliminate the plaguing problem of scalability for such systems due to the stringent requirements. Despite there being wide applications that employ inanimate objects as markers for interaction, human body parts as primary markers in applications are still an obscurity. Applications that are mentioned and implemented above, justify the readiness for research and development into spatial computing environments. Spatial computing has great use in the forthcoming future and more widespread application is anticipated as technology in 5G communication and cheaper computing power become readily available, combining real and virtual environments, to give rise to the next generation of computing environments.

REFERENCES

[1] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios, Protopapadakis, ”Deep Learning for Computer Vision: A Brief Review”, Computational Intelligence and Neuroscience, vol. 2018, Article ID 7068349, 13 pages, 2018.

[2] Shekhar, S., Feiner, S.K. and Aref, W.G., 2015. Spatial computing. Communications of the ACM, 59(1), pp.72-81.

[3] Lu, W., Tong, Z. and Chu, J., 2016. Dynamic hand gesture recognition with leap motion controller. IEEE Signal Processing Letters, 23(9), pp.1188-1192.

[4] J. Dudley, H. Benko, D. Wigdor and P. O. Kristensson, ”Performance Envelopes of Virtual Keyboard Text Input Strategies in Virtual Reality,” 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2019, pp. 289-300.

[5] Y. Zhang, W. Yan and A. Narayanan, ”A virtual keyboard implementation based on finger recognition,” 2017 International Conference on Image and Vision Computing New Zealand (IVCNZ), 2017, pp. 1-6.

[6] Pulli, K., Baksheev, A., Kornyakov, K. and Eruhimov, V., 2012. Realtime computer vision with OpenCV. Communications of the ACM, 55(6), pp.61-69.

[7] Erhan, D., Szegedy, C., Toshev, A. and Anguelov, D., 2014. Scalable object detection using deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2147-2154).

[8] G. Chandan, A. Jain, H. Jain and Mohana, ”Real Time Object Detection and Tracking Using Deep Learning and OpenCV,” 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 1305-1308, doi: 10.1109/ICIRCA.2018.8597266.

[9] Rahim MA, Shin J. Hand Movement Activity-Based Character Input System on a Virtual Keyboard. Electronics. 2020; 9(5):774.

[10] Wu, Y., Lim, J. and Yang, M.H., 2013. Online object tracking: A benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2411-2418).

[11] Ming-I Brandon Lin, Ruei-Hong Hong Yu-Ping Huang (2020) Influence of virtual keyboard design and usage posture on typing performance and muscle activity during tablet interaction, Ergonomics, 63:10, 1312-1328

[12] Suwen Zhu, Tianyao Luo, Xiaojun Bi, and Shumin Zhai. 2018. Typing on an Invisible Keyboard. ¡i¿Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems¡/i¿. Association for Computing Machinery, New York, NY, USA, Paper 439, 1–13.

[13] Hangu¨n, B. and Eyeciog˘lu, O¨ ., 2017. Performance comparison between OpenCV built in CPU and GPU functions on image processing operations. International Journal of Engineering Science and Application, 1(2), pp.34-41.

[14] Lee TH., Lee HJ. (2018) A New Virtual Keyboard with Finger Gesture Recognition for AR/VR Devices. In: Kurosu M. (eds) Human-Computer Interaction. Interaction Technologies. HCI 2018. Lecture Notes in Computer Science, vol 10903. Springer, Cham

[15] Jitendra Madarkar and Poonam Sharma, ”Occluded face recognition using noncoherent dictionary”, in Journal of Intelligent Fuzzy Systems, vol , pp, 2020.

[16] Jitendra Madarkar and poonam Sharma, ”sparse Representation for Face Recognition: A Review Paper”, IET Image Processing, vol.-15, pp-1825- 1844, 2021.

[17] Jitendra Madarkar and Poonam Sharma, ”Sparse Representation based Face Recognition using VGGFace”,International Conference on Machine Learning and Big Data Analytics (ICMLBDA),2021.

★★★