International Journal of Modern Agriculture, Volume 10, No.2, 2021
ISSN: 2305-7246

Smart Video Surveillance of Human Interactions in Crowded and Strongly Occluded Scenarios Using Dense Trajectories from Video and Wearable Cameras

By Mrs. Sabitha P[1], Mahadev K[2], Naveen Ramanathan[3], Hariharan[4]

[1]Asst. Professor in CSE Department, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India
[2, 3, 4] UG Scholars, CSE Department, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India

Abstract— The detection and recognition of conversational body gestures while in group discussions and conferences, and in crowded situations is the foremost concept we are addressing. Many related detections and recognition systems have failed to work during dense situations, which has multiple challenges such like cross-contamination among witnesses, heavy occlusions and heavily dynamic backgrounds that are influenced by the condition of the witnesses, light situations or obscurations and shadows. This makes the detection further complicated to analyze with the existing techniques of computer vision. We are approaching this problem by fusing multiple modal resources that is, by using video and body-worn cameras where the data will be continuously recorded. By using video modality, we analyze RGB-D for individual hand and arms trajectory tracking and gesture trajectory recognition concerned with the depth video and synchronized colour. This gesture detecting method uses motion, targeted witness, and depth-based particle filters that will emphasize the efficiency and accuracy largely, in the context of the witness performing gestures toward the camera device within a crowded environment and in front of moving loud and disarrayed environments. We are using the wearable camera to emphasize the natural association between communication, expressions and gesticulations during conversations, which is more efficient than the video. Our proposed approach fuses both the determinations from classifiers from the CCTV video source and body wearable camera modalities, we can increase the area under ROC performance largely. Also, the huge occlusion under the video source is remunerated by the data from the wearables. Therefore, we applied our approach to detect and recognize the speaking state, improving the solid connection found in the discussion between hand gesticulations and conversations**.**

Index terms — Gestures, RGB-D, MILES, Trajectory, Wearable Camera.

I. INTRODUCTION

Surveilling people in strongly crowded situations, especially at events, becomes a challenge for any health and safety executive as the safety precautions they have implemented will be evaluated to acknowledge how effectively their protocols were executed. There are various considerations to be included to design these safety protocols.

Firstly, the executives need to monitor the total number of people for ensuring that the threshold capacity is not exceeded. Secondly, they must be identifying potential crowd problems such as preventing the intensification of public disturbance. Finally surveilling the population that helps in preventing the local occluding. The most important factor for surveilling under such circumstances is setting up surveilling devices in areas that are highly occluded. The areas that are likely to monitor are entrances and exits, areas where people queue, popular stalls and, refreshments and confined spaces. Major drawbacks if the surveillance is done by staffs manually, can cause them distress and anxiety. When there is a huge unexpected crowding, the staff might consider having additional staff thus decreases their workloads. The only positive with staff surveilling within the crowd is that they can observe people's body language and identify the signs of interactivity and engagement in that atmosphere. Executives on the ground can also help diffuse situations when there is turbulent behavior within the crowd.

Computer vision and big data applications are overwhelming most of the scope in the industry and research field. Amongst the popular examples of Computer vision, the role of detecting human activities through their gestures and body languages is a huge contributing factor not only in the field of surveillance but also in any other field that includes the prediction or analysis of human activities. Similarly, the role of big data that is collected through video streams from CCTV cameras is evenly significant as other sources like the data from agriculture, sensors, medical, social forums resulted from IoT devices or space experimentation. Surveillance sources have a larger offering to unstructured big data. Surveilling cameras are implemented in all areas where safety has many effects. As mentioned earlier manual surveillance will result in tedious and time-consuming. Security can be described in various terms in different circumstances like fraud identification, disorder detection and other uncertainties in strongly occluded scenarios. The exceptional or irregular movement interpretation in a strongly occluded and crowded video is very complicated due to several real-world confinements.

Most of the works concerning the recognition and detection of gestures focuses on the issues or situations where there is a perfect view of the person performing a symbolic gesture, usually from frontside view. For certain researches, the method is quite related, gesture detection and pose estimation that is use of its skeleton. We are solving this problem by fusing multiple modal resources that is, by using video and body-worn cameras where the data will be continuously recorded. For the video modality, we are using RGB-D for people’s hand trajectory detection and gesture trajectory recognition based on detected colour and depth from the video. This gesture detecting method uses body motion, prominent skin and depth-based particle filters that are capable of improving the tracking accuracy considerably, in the context of the beholder performing the gesture toward the camera device within a crowded environment and in front of moving loud and disarrayed environments.

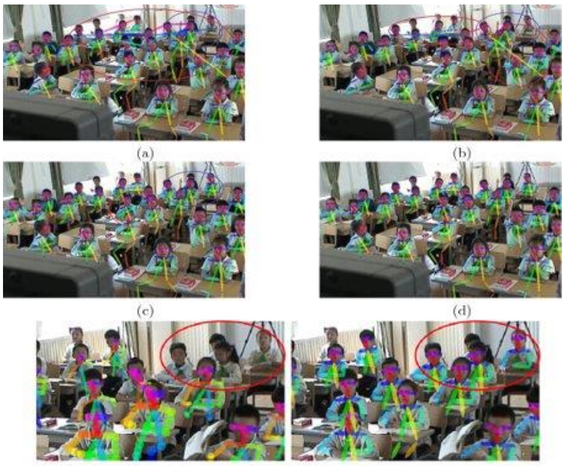


Fig.1.a. Analysis of the learning states of students using their conversational gesture.

Fig.1.b. Hand gesture recognition using Kinect

Fig.1.c. Wearable camera device worn by the members.

We are utilizing the body worn camera to use the normal connection among discourse and signals during discussions, which is more effective than the video. At first, we are tending to the recognition of conversational hand motion utilizing a dataset gathered during a genuine blend occasion with solid relational impediments from which we break down what thick directions in existence are more representative towards a signal in the source of video. We join wearable camera and video data source in a choice level way utilizing the complementarity among modalities, especially for situations where impediments in the video data source are too solid to even think about having a reasonable perspective on the individual playing out the signal, appearing upgrades over unimodal methodologies.

We extricate trajectories utilizing the technique for thick directions which have demonstrated to be a proficient portrayal for human movement acknowledgment. We at that point investigate the effect of uproarious information that is the solid impediments of the members on the general execution, both static and powerfully on schedule; lastly, we utilize our strategy to distinguish talking status, utilizing the connection between individuals' motions and discourse. To distinguish talking status, we utilize the names for the talking status of all members gave by the dataset. Some huge impediment from the source video is repaid by the data from the body wearables. We implemented our methodology to identify and perceive the talking state, improving the strong relationship discovered in the conversation between hand signals also, discourse. Also, deals with boisterous conversational gatherings have determined that body wearable detecting choices can give extra data when acquiring with video source. Along these lines, each methodology can give diverse data to comprehend the occasion, depending on one methodology when the additional are abstaining or utilizing their complementarity.

II. BACKGROUND AND RELATED WORK

The current frameworks are centered around situations where the individual is just performing their own and natural motions that are obvious. Practically many existing works on the discovery and ID of signals center around situations where ever there is any reasonable perspective on the body conveying an emblematic motion, for the most part from the front side. Numerous datasets have given more than a million recordings of each individual in turn performing communication through signing motions before a Kinect (Fig 1.b) [1] [2], either trimmed or with constant signals. Shockingly, this doesn't estimate most of the genuine circumstances where signals are utilized. There are numerous issues we face, in actuality, circumstances like cross-pollution between subjects, solid impediments that make it difficult to see the subjects now and again solid changes in appearance for a similar subject and non-fixed foundations, which are influenced by the situation of the subjects, lighting conditions or shadows.

Starting with the research done by P Wang and the team in 2016 [3] on large-scale continuous gesture recognition with the help of deep learning's convolutional neural networks, which has given many ideas for adapting for our research. The main idea illustrated by P. Wang was determining the beginning and the end frames for each and every gesture utilizing the Quantity of Movement. The major ideology of this research is that, by using the Quantity of Movement with assumptions, that all gestures start from a clear and similar posture will help in predicting the gesture with high accuracy. But the major drawback of this research is

that, in real-life situations like cross-contamination between subjects, strong occlusions where gestures with high accuracy cannot be predicted by starting with a similar and clear pose. Another research was done by Chai and Liu 2016 [4], which inherited the Wavelet and Fourier transform to recognize and extract conversational gestures, contributed to our research on understanding how two streams recurrent neural networks can help continuous large-scale gesture recognition. We also adapted the idea of their research's implementation that was based upon dataset that has a rather perfect side-view of the witness and only thirty seconds video.

The ChaLearn by Yibing Zhao and Shuai Zhou 2016 [5] has provided over 40000 videos of an individual acting body and sign language gestures in front of a Kinect Device (see Figure 1()), each cropped or with continuous gesticulations and poses. Several researchers have used this dataset to approach the difficulty of gesture recognition under these restrictions and occluded situations. There was another fascinating research by N. Camgoz, S. Hadfield 2016 [6] that also helped us for adaptation, where they have been trained as an end-to-end deep network utilized for simultaneous and continuous gesticulation recognition by combining the understanding of both the feature representation and the classifier. That network mainly performs the three-dimensional Space-time convolutions for extracting the features that are associated to all the presentation and movement from all the masses of colour frames.

One major contributing work is done by Quek and McNeill 2002 [7] on Computer-Human Interaction which was a gateway for many HCI applications and its expansion. This is the first work that used Multiple Instances Learning for detecting gestures, in other words, usage of Multiple instances learning for general human activity recognition. The most resembling our research concerning the implementation of Multiple Instance Learning (MIL) for recognizing gestures that were submitted by Ali and Shah [8] showed techniques based on multiple instances for common action recognitions. The major drawback of this research is that it did not include any hand and arms conversational gestures as a part of their models. Their datasets were taken from the movies, YouTube and video hosting, sharing, and services platforms.

To our knowledge, there are very few researches that have contributed to human gesture recognition in strongly occluded places. But we are very thankful for many works that helped us in bricking up better human activity recognition in strongly occluded scenarios.

III. TAXONOMY

A. Strongly Occluded Situation

We are utilizing the MatchNMingle dataset which is a multimodal asset used for the examination of social associations in the strongly crowded situations. This dataset gives data to up to 70 individuals while blending uninhibitedly for 30 minutes. During one of three distinctive day occasions, members were essential for a speed date occasion, each followed by a blended meeting.

Meanwhile we center around the identification of motions in the occlusion during standing multi-party discussions, we will just utilize the blend part of MatchNMingle Dataset yet we theorize that our experiences here can be additionally applied to a situated situation with respect to the speed meetings. For the blend meeting, the members were not educated at all, and they can move openly through the blend zone or leave it freely. They can likewise arrange food or beverages during the whole occasion.

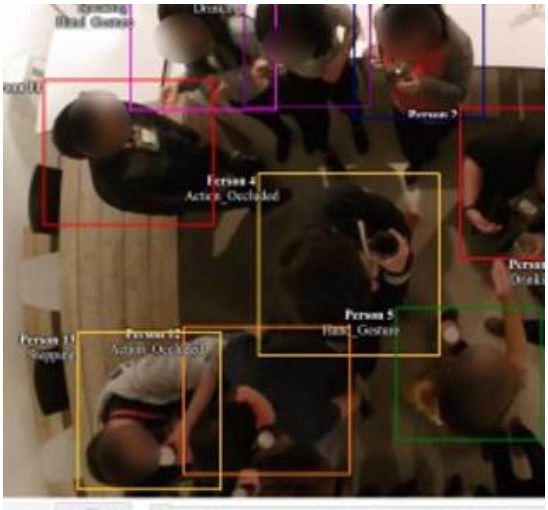


Fig.2. Analysis of conversational gesture in a strongly occluded venue.

Consequently, their signals are inalienable with respect to the social collaborations they are having with different members or with individuals from the staff. Every member wore a shrewd identification stuck around the neck recording triaxial speed increase at 20 Hertz meanwhile the whole occasion. Additionally, the video was recorded at 20 FPS from a higher place. At last, the dataset likewise gives the manual comments of the social activities for all members and the fundamental truth for their situations in the picture.

B. MatchNMingle Dataset

MatchNMingle is a multimodal dataset utilized for the examination of conversational groups and exhibitions happening with the extreme crowd. MatchNMingle emphasizes the efficiency and notable accuracy of wearable camera devices and CCTV cameras that are fixed at the hall or room's corners to record the human interactions of people during real-life speed conversations which were then followed by a party.

This dataset is open source and is widely available for studies and research purposes concerning the human activity recognition system. The dataset is under a license for it to be used in any commercial projects. As illustrated in the figures, the dataset covers the activities performed by a group of people. It also covered their emotions and engagements, which in other words called sentiment analysis. But it differs from a regular sentiment analysis as this data set has not only the frontal photos but also the survey responses which was been asked to be filled by the participants.

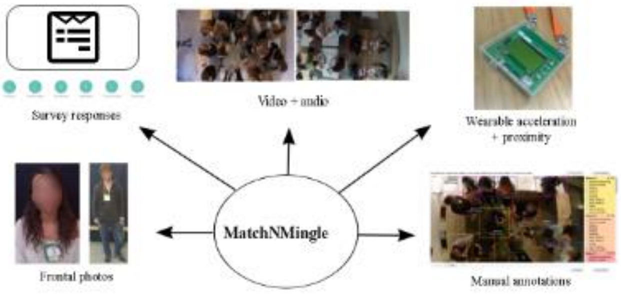


Fig.3. The MatchNMingle dataset which is openly available for research purposes, under an End-User License Agreement (EULA)
https://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/ **[9]**

The Figure 3 illustrates the dataset that has the biggest number of members, most continued recording time and the most huge set of standard commentaries used in terms of social actions are available in this setting in a natural life situation. It is consisted of data obtained from a wearable camera and paired concurrence which was happening for hours, audio, character surveys, importantly video, interacting with social media and human responses in public forums. Members' locations and gathering formations were annotated in a manual process as were social actions such as speaking and hand gesture for 30 minutes at 20 FPS addressing it the first dataset to consolidate the commentary of such ideas in this connection.

C. Multiple-Instance Learning via Embedded Instance Selection (MILES)

Multiple instance complications resulting from the circumstances where the training class labels are connected to the sets of samples which are termed as bags and rather than unique samples within each bag which are called instances. Earlier Multiple Instances Learning algorithms and representations were designed and formed on the premise that every individual bag will be positive and accurate only if one of its instances is also positive. However, the hypothesis works fine in a drug activity prognostication and foresight problem and is restrictive for other utilization, especially those in the computer vision space.

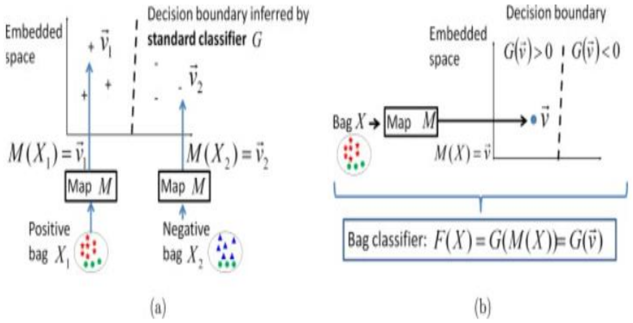


Fig.3. Illustration of Multiple instances Learning's classification of the ES (Embedded Space) paradigm: training (a) and test (b) **[10]**.

Multiple Instances Learning through Embedded Instance Selection also abbreviated as MILES is used in the conversion of the multiple instances learning predicament to a standard supervised learning predicament that does not require the hypothesis comparing between the instance labels and the bag labels. MILES is used in the mapping of individual bags into feature space determined by those instances present from the training bags through an instance relationship measure. This feature routing often provides a huge number of excessive or unrelated features.

Fig.5. Clustering process in time and space to develop bags of thick trajectories obtained from the video source

IV. PROPOSED APPROACH

We are taking care of this issue by melding various modular assets that are, by utilizing video and body-worn cameras where the information will be persistently recorded. By utilizing video methodology, we are utilizing RGB-D for human hand direction following and motion direction acknowledgment dependent on synchronized tone and profundity video. This signal following technique utilizes striking skin, movement, and profundity-based molecule channels that are equipped for improving the following exactness significantly, with regards to the observer playing out the motion toward the camera gadget and before moving, loud and jumbled foundations. We are utilizing the wearable camera to use the characteristic connection among discourse and motions during discussions, which is more productive than the video. At first, we are tending to the discovery of conversational hand motion utilizing a dataset gathered during a genuine blend occasion with solid relational impediments. investigate which thick directions in existence, are more delegate for a signal in the video.

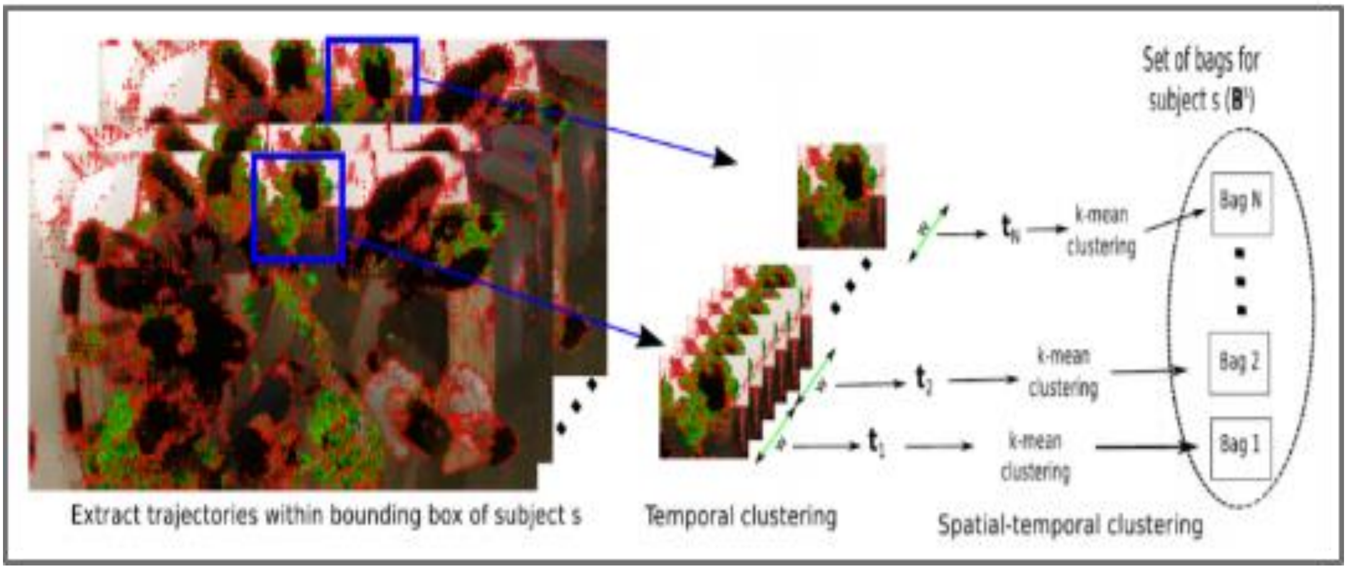
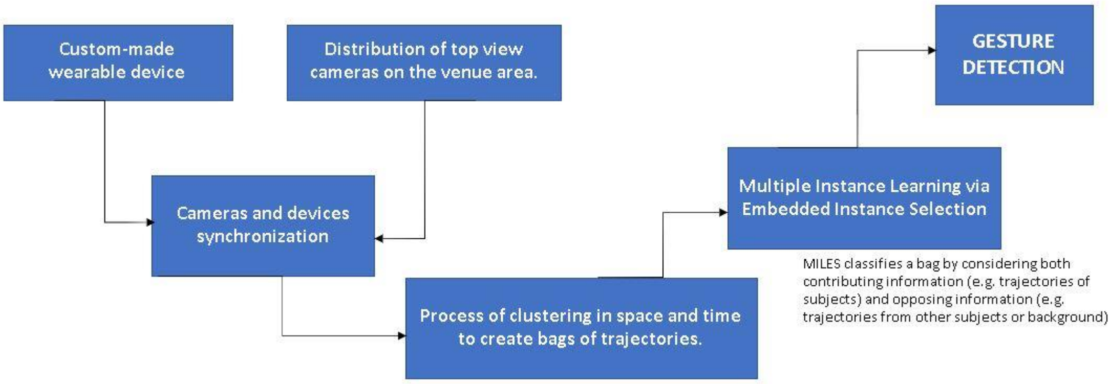


Fig.6. Architecture Diagram

We join the body wearable camera source and the video source by speeding up in a choice-level way utilizing the complementarity and complexity between the modalities, especially for situations when and where impediments in the video will be too solid to even think about having a reasonable perspective on the individual playing out the


Custom-made wearable device
Distribution of top view cameras on the venue area.
Cameras and devices synchronization
Process of clustering in space and time to create bags of trajectories.
Multiple Instance Learning via Embedded Instance Selection
GESTURE DETECTION
MILES classifies a bag by considering both contributing information (e.g. trajectories of subjects) and opposing information (e.g. trajectories from other subjects or background)

motion, showing upgrades over unimodal methodologies.

A. Dense Trajectories Extraction from Video Source

Initially there is a need to extract trajectories using the dense trajectories extraction methodologies proposed by Wang and team. They are very accurate and efficient presentation for activity recognition [11] [12] [13]. The dense trajectories are imparted to cover the 20 frames with a length L to form a complete frame.

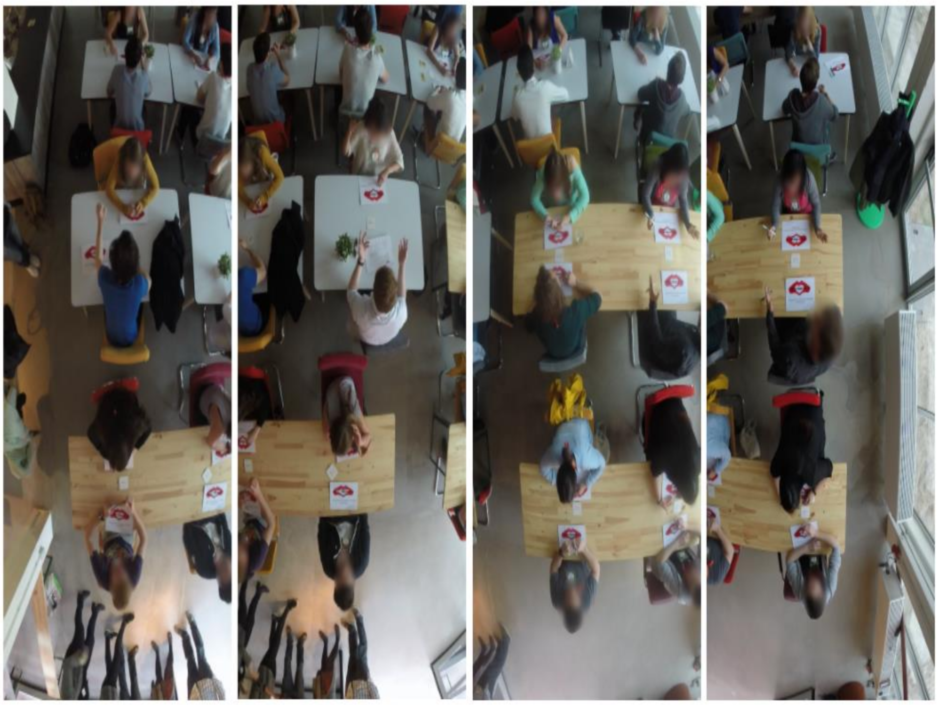


Fig.7. Screenshots from the Video Source.

Utilizing the jumping boxes for every member on each the outline we make a voxel following the member after some time. In this way, we lessen the amount of trajectories for those which are around or from every member by choosing just those inside this voxel. This choice likewise represents trajectories that will begin outside of the voxel however we will enter it and those that start inside the voxel and float out. For bounding box extraction one can utilize any current apparatus for this reason [14] [15] notwithstanding, we utilize the ground truth explanations to stay away from additional tainting. Our technique utilizes packs of video highlights, so it ought to be vigorous against little moves of the jumping boxes.

In any case, we leave any investigation of the effect of the blunders in the identification and following of individuals on the general gesture location for future and advanced studies and research this will stand outside of the extent of this research. Please note, because of the crowdedness and congestion of the scene, jumping boxes of various subjects can vigorously cover. The bounding box can subsequently contain the trajectories of both witness's interest and 'foundation' subjects. Luckily, our various occurrence learning MILES approach will be helpful in representing the feature of duality.

B. Multiple Instances Learning (MIL) Utilization

We are using Multiple Instances Learning through Embedded Instance space selection which can help in classifying each bags by considering all the supporting data such as trajectories of witness s (let's assume) and engaging data such as trajectories from other backgrounds or objects or subjects. It does so by generating a notion in an embedded space and connecting all the instances to this thought.

Therefore, by not considering other Multiple Instances Learning approaches wherever any least +ve instance from any individual bag can automatically transform it into any positive bag, wherein MILES does not hold this confinement. This helps in allowing us to evaluate the role of specific instances in the analysis of each bag in the collection. To understand how MILES helps us in mapping each bad into feature space that are absolutely defined by instances within the training set with the help of their similarities.

To understand this more specifically, let us assume a set of bags concerning our all witnesses who are participants as B = {B^1, B^1, B^1,..,B^S} and assume B_a be an individual bag belonging the set B, where a is {1,2,3…,N} and N is the summation of absolute number of bags for S number of subjects, nothing but witnesses. We can say that any bag B_a has the value of similarity between the current one and rest of the instances belonging to the training can be measured by

$$s(\mathbf{x}^k, \mathbf{B}_a) = \max_b \exp\left(-\frac{||\mathbf{x}_{ab} - \mathbf{x}^k||^2}{\sigma^2}\right)$$

The left hand side value will be the calculation of similarity, and can be identified by nearest instance from the bag till end of the concept. By understanding our approach we can say that any individual bag will possibly be embedded into space having the coordinates m(B_a) can be evaluated by the following formula.

$$\mathbf{m}(\mathbf{B}_a) = [s(\mathbf{x}^1, \mathbf{B}_a), s(\mathbf{x}^2, \mathbf{B}_a), ..., s(\mathbf{x}^{n_a}, \mathbf{B}_a)]^T$$

Here the n_a will be the summation of instances belonging to the major training set. The above formula will output data frame of bags under training in the ES.

$$\mathbf{m}(\bar{\mathbf{B}}) = [\mathbf{m}(\mathbf{B}_1), ..., \mathbf{m}(\mathbf{B}_A)]$$

The formula for representing the classification of all the new bags can be expressed by the following formula.

$$y = \text{sign}(\sum_{k \in I} w_k^* s(\mathbf{x}^k, \mathbf{B}_{new}) + b^*)$$

And I being the subset of instances,

$$(I = \{k : |w_k^*| > 0\}).$$

To understand more about the MILES algorithm do visit and review [16]. Our last expression to determine the relationship between x_{ij} where (x) is the instances for the classification for a new bag $B_{i,}$ where J = 1…N_i (Summation of instances within the bag).

C. The contribution of the wearable camera device to improve the accuracy of recognition

A personal wearable camera device that is worn individually records the triaxial-acceleration up to 20 Hertz. For each member, there is a need to evaluate the extent of the acceleration, that will output in four different time series, that we can impart features with the help of the sliding window technique. We approach those actions where the direction is significant, with the help of triaxial-time -series.

$$\mathrm{Mod}(A) = \surd\ (x^2 + y^2 + z^2)$$

At that point, for every one of similar sliding-windows as characterized in the past segment, we remove includes that have demonstrated to be productive to investigate human activities from wearable speed increase. These highlights are principally factual and ghostly, where the measurable highlights zeroed in on mean and fluctuations from every hub and the greatness, and otherworldly utilizing the force unearthly thickness. All highlights are linked to acquire a component vector for each window and afterward arranged utilizing calculated relapse.

D. The fusion of both the sources

For combination, we chose a choice-level joining approach. In this way, the surmised back likelihood of the video data obtained from CCTV and the body wearable classifiers that are utilized for a contribution for a third classifier. We pick choice combination rather than early combination (for example link highlights) as we intend to keep a steady and reasonable component space. This is giving an understanding that the MILES can plan each sack to an installed space characterized by its occurrences likenesses, this cannot be implemented for highlighting from the wearable speed increase, subsequently making early fusion and combination impractical.

V. DISCUSSION

A. Benefaction from Multiple Instance Learning via Embedded Space

We can understand that largest of the trajectories are determined by the Multiple Instance Learning via Embedded Space (MILES) as of huge participation, are that corresponds to arms and hands. Here implements even if different areas of the human body for the witness, other witness or the dynamic background which is also under motion. Therefore, our MILES method succeeds up to a specific level and that is, its intention to remunerate as cross-contamination of the video data. Nevertheless, we could also conclude from the illustrations that the failure situations where the trajectories that correspond to another witness, cross-contamination, will be given huge benefaction. Concerning the case, specially particular, was exceptional, meanwhile, those divisions the two witness were involved in a conference and the witness prompting some cross-contaminations were also gesticulating. Therefore, the Multiple Instance Learning via Embedded Space learns precisely that those trajectories were resembling to a gesticulation, though fails to distinguish the witness.

B. Improving the detection of the gesture while the speaker's status modulates

At last, we examine the contrasts between the outcomes found for gesture identification and talk status recognition. This test attempts to use the common connection among discourse and gestures during discussions. Outputs represent that for the speaker's conversation status identification the wearable camera is more efficient comparing with the video resulted from CCTV devices and the qualities were like those found in past work utilizing the equivalent dataset. Regardless conversely with the gesticulation identification, the outcomes for video data that is both standard and MILES will be extensively more inferior than those observed for gesticulation location and the combination doesn't develop covering the unimodal methodologies.

Every individual recurrence circulation activity can improve the understanding of why our main procedure is perhaps frequently working for the conversing situation contrasted with the gesticulation identification. Us, tracked down that approaching normal members went through 400+ seconds talking and 300+ gesturing. However, two activities cover approximately 200 seconds. This investigation shows that not every one of the gestures is identified with discourse and not all discourse was carefully joined by the gesture. Consequently, our strategy could attempt to decipher a gesture consistently as a feature of the discussion which isn't the situation. Besides as discourse is not joined by the gesticulation, we may have the large number of bogus contradictions concerning minutes among just discourse. This can likewise clarify why our body-worn camera device will be more exact and efficient in distinguishing the talking status comparing with video data or combination as the gadgets sense development chiefly from the middle that is bound to vacate when individual talks and not concerning hands.

VI. EXPERIMENT

Our results of Gesture Detection utilizing Unimodal classifier with respect to their fusion is given in the following.

MEAN AUC (± DEVIATION)	Video Source (CCTV)	Body Wearable Camera	Multiple Instance Learning via ES
AUC-ROC	**0.61 ± 0.08**	**0.7 ± 0.06**	**0.70 ± 0.12**

Our results for the influence of complex in gesture recognition, and the mean area under curve using test and train subsets which is completely disputed is given in the following.

TEST	TRAIN	MEAN AUC (± DEVIATION)
Noisy	Clean	0.68 ± 0.10
Clean	Clean	0.70 ± 0.08
Noisy	Noisy	0.69 ± 0.11
Clean	Noisy	0.71 ± 0.09

Our results of Speak Detection utilizing Unimodal classifier with respect to their fusion is given in the following.

MEAN AUC (± DEVIATION)	Video Source (CCTV)	Body Wearable Camera	Multiple Instance Learning via ES
AUC-ROC	**0.5 ± 0.05**	**0.7 ± 0.10**	**0.66 ± 0.15**

Our test resulted in a mean AUC-ROC of 0 66 ± 0 15. Thus as we theorized the utilization of complementarity among modalities expand the presentation of the identification while contrasted with the unimodal methodologies 0 68 and 0 64 for video and the body wearable camera individually.

VII. CONCLUSION

Concerning the conclusion, we manifested our approach to recognize gestures during congested and strongly occluded situations with the help of bags of thick trajectories obtained from the video source and the wearable camera source. This disclosure is especially complicated for blend situations as all present large subject's cross-contamination, and heavy occlusions amongst additional difficulties.

The most complicated thing to succeed is the extremely noisy and clamorous video source, where we have implemented the Multiple Instances Learning via Embedded Space (MILES) method, which confirmed to be capable to manage difficulties such as noisy and dynamic backgrounds and environments, also cross-contamination amongst subjects until a particular point. Furthermore, we examined the participation of instances within the classifier determining that, it learns from trajectories expressing a witness's gesturing and neglects those from the environment.

We likewise explored the effect on the presentation of uproarious information because of subject cross tainting and impediments both static and dynamic on schedule way. This investigation showed that intertwining modalities additionally makes up for those minutes where the certainty of MILES classifiers rots because of impediments. At long last we applied our strategy to recognize twofold talking status utilizing the reason that gestures and discourse are by and large interlaced.

REFERENCES

[1] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," IEEE Trans. on Multimedia, 2013.

[2] C. Wang, Z. Liu, and S. Chan, "Superpixel-based hand gesture recognition with kinect depth camera," IEEE Trans. on Multimedia, 2015.

[3] P. Wang, W. Li, Z. Gao, C. Tang, and P. Obunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," IEEE Trans. on Multimedia, 2018.

[4] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," Intern. Conf. on Pattern Recognition (ICPR), 2016.

[5] J. Wan, S. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016.

[6] N. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Using convolutional 3d neural networks for user-independent continuous gesture recognition," Intern. Conf. on Pattern Recognition (ICPR), 2016.

[7] F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. McCullough, and R. Ansari "Multimodal human discourse: Gesture and speech," ACM Trans. on Computer-Human Interaction, 2002.

[8] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2010

[9] S. KanagaSubaRaja, S. Usha Kiruthika, (2015) 'An Energy Efficient Method for Secure and Reliable Data Transmission in Wireless Body Area Networks Using RelAODV', International Journal of Wireless Personal Communications, ISSN 0929-6212, Volume 83, N0. 4, pp. 2975-2997.

[10] S JaumeAmores, "Multiple instance classification: Review, taxonomy and comparative study", Computer Vision Center, Computer Science Department, UAB, Spain, 2013.

[11] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action Recognition by ¨ Dense Trajectories," Colorado Springs, United States, pp. 3169–3176, Jun. 2011. [Online]. Available: http://hal.inria.fr/inria-00583818/en

[12] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," Intern. Journal Computer Vision, 2013

[13] J. C. van Gemert, M. Jain, E. Gati, and C. Snoek, "Apt: Action localization proposals from dense trajectories," British Machine Vision Conf. (BMVC), 2015.

[14] R. Stewart, M. Andriluka, and A. Ng, "End-to-end people detection in crowded scenes," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016

[16] Y. Chen, J. Bi, and J. Wang, "Miles: Multiple-instance learning via embedded instance selection," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2006.

[17] P. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," Multimedia Systems, 2010.

[18] Murugan, S., Jayarajan, P., &amp; Sivasankaran, V. Majority Voting based Hybrid Ensemble Classification Approach for Predicting Parking Availability in Smart City based on IoT.

[19] Efficient Contourlet Transformation Technique for Despeckling of Polarimetric SyntheticApertureRadarImage Robbi Rahim, S. Murugan, R. Manikandan, and AmbeshwarKumarJ. Comput. Theor. Nanosci. 18, 1312–1320 (2021)