ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse68912020.pdf
https://doi.org/10.30534/ijatcse/2020/68912020

The Geometrical Based Lip-Reading Techniques of Multi-Dimensional Dynamic Time Warping MDTW and Hidden Markov Models HMMs in the Audio Visual Speech Recognition

Muhammad Ismail Mohmand[1], Amiya Bhaumik[2], Poom Kumam[3], Qayyum Shah[4], Muhammad Humayun[5]
[1, 2]Faculty of Engineering at Lincoln University College LUC, Wisma Lincoln, Petaling Jaya Selangor Darul Ehsan, Malaysia. ismail.mohmand@lincoln.edu.my, and amiya@lincoln.edu.my
[3]Professor in the Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand. poom.kum@kmutt.ac.th and poom.kumam@mail.kmutt.ac.th
[4, 5]Department of Basic Sciences & Islamyat U.E.T, Peshawar, Pakistan. humayunchemist@uetpeshawar.edu.pk and qshah08@gmail.com

ABSTRACT

This paper portrays a programmed lip perusing framework comprising of the two primary modules such as, a pre-preparing module ready to separate lip-geometrical data from video grouping and a characterization module to recognize visual discourse dependent on unique lip-developments. The recognitions execution of the planned framework has been evaluated in the acknowledgment of English digits staring from 0 to 9 spoken by the speaker's in the video groupings accessible in Clemson University Audio Visual Experiments (CUAVE) database techniques. The extractions of lip-geometrical features was completed utilizing a blend of skin-shading channel, an outskirt following calculation and an arched Hull-approach as well as proposed strategy was contrasted and the well-known 'snake' procedure and was found to expand the lip-shape extractions execution for database are considered. The lip-geometrical features including stature, width, proportion, territory, edge as well as different mixes of features were assessed to the figure out which plays by the speaking to discourse in the visual area in the use of three discrete arrangement techniques, in particular optical stream, Dynamic Time Warping (DTW) and another methodology named Multi-dimensional DTW and Hidden Markov Model (HMM). The experiments shows that the proposed framework is fit for an acknowledgment execution of 74% simply utilizing lip stature, conventional appearance-based Discrete Cosine Transform DCT techniques of the lip-width and proportion of these features exhibiting that framework can possibly be fused in a multimodal discourse recognitions framework for the use of energetic environments.

Key words: Lip reading techniques, skin color, convex filtration, multi-dimensional DTW and HMM.

1. INTRODUCTION

In the recent developments of automatic speech recognition ASR system by using the audio only method is capable to realize a standard of execution reasonable for some pragmatic applications techniques. Nonetheless, within the sight of nosy sound clamor, for example, that initiate in air terminals as well as train stations, either in closeness of different speakers in workplaces or at parties, the presentation of such frameworks is known to turn out to be seriously corrupted. Various examinations have uncovered that the data contained in discourse sign is firmly identified with that found in lip moments and if data in regards to the last is incorporated the observation execution of the two people and machines can be improved. In uproarious situations, people can diminish discourse acknowledgment blunders by utilizing the speakers lip moments [1] and to be sure numerous individuals with hearing challenges depend on lip perusing to give most of the discourse data they get. There are two fundamental issues that should be tended to in planning and actualizing a lip-reading framework and the first decision of visual features while the second to the improvement of an incredible procedure for the features extractions from the video streams. Different choices of the visual features extractions have been proposed in composition starting late and these can be assembled into three essential arrangements such as appearance-based features extractions, the shape-based features extractions and the incorporation of both appearance based features extractions and the shape based-features systems [2].

A couple of strategies have-been planned in the previous for the lip-understanding techniques. In [3] sis sorts of the visual-component were taken a gander at the discrete cosine transforms DCT has been investigated to the preeminent execution and it's connection with discrete wavelet transforms DWT with the strategy of the principal component analysis PCA as well as active appearance-model AAM moves close to the lip geometrical techniques. In [4] lip-perusing approaches such as motion history imaging MHI have been depicted in the progression improvements were gotten and total subtraction systems gave an amazing depiction of the visual talk remembers for a single grayscale picture and the artificial neural network ANN, were used to amass MHI pictures. In any case, this strategy was viewed as into the single picture. Therefore in [5], a methodology be situated to the delineated that depicted lip-scrutinizing by taking care of its optical stream with quantifiable of the optical stream piece creature as well as the utilization to the shape section vector for the setting up with machining classifier. While on the downside of this system to optical stream strategies are fragile as well as turn of photographs further down investigation and the composing classifications proposes of appearance-based frameworks by and large produce favored talk confirmation results over shape based methodologies as the past pass on more data and the execution of the last is of an extra sales of trouble in perspective on the need to get a precise division of the lips in giving unfathomable geometrical-highlights [6]. In any case, the similar unwavering nature of the methodology to show up based features is for the most part lower because of its inexorably significant affectability to condition, for example, illumination as well as head present [7]. So in the inspiration of present effort is in and that manner to build up a technique that produces lip-features of critical worth like that ordinary of the appearance based frameworks. However that can be gotten fearless idea of the shape based procedures in multi-dimensional dynamic time warping MDTW Techniques.

The basic steam of this novel paper is to approaching to the lip-geometrical system briefly to investigate the multi-dimensional dynamic time warping MDTW and Hidden Markov Model HMM whether geometrical information procured by the shape-based examination. In this paper shows another procedure which incorporates at first extraction of the lip-geometrical as well as besides gathering of highlights, and the lip geometrical are gotten by using the skin concealing channel, the affirmation of the raised body. In this manner, the lip-shape highlights are evacuated and the range of the verbalization as well as gathering to the practiced approach named MDTW and Hidden Markov Model HMM in the broad media discourse acknowledgment framework and the suitability of the strategy are assessed by using CUAVE database techniques [8].

2. SKIN COLOR FILTERATION TECHNIQUES

The work portrayed in this paper ware creating and utilizing the Microsoft Visual C# 2015 as well as used the open sources picture preparing library such as Open-CV. The parts of new approaches for lip-geometrical features extractions are appeared in Figure 1 respectively. The speakers pictures gained from video documents of CUAVE database are edited to mouth district utilizing in face identification procedure pursued by a mouth location process. The skin location procedures are used to fragment the lip as well as non-lip territories in the mouth locale. At long last, the lip shape features, stature, width, proportion, zone and edge, are extricated by applying fringe following and the arched structure technique.

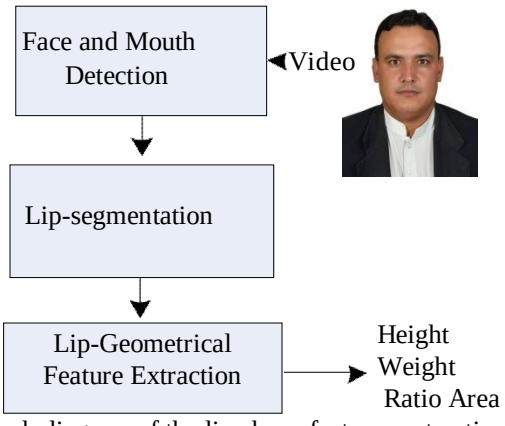

Figure 1: The block diagram of the lip-shape features extractions techniques.

2.1 Face and Mouth Detection Technique

An extensive scope of procedures are accessible for the extricating face and mouth region for example, layout coordinating and movement location and that strategy received to initially utilized in [9] that structures a results of the each picture area and at a few scales and afterward to utilize the outcomes to prepare a powerless classifier utilizing Ada-Boost while single solid article classifier would be able to be framed by falling these feeble classifier. Therefore the benefit of having frail classifiers working in early preparing can disengage areas of the article areas, in this manner enabling more noteworthy groupings of exertion to be applied as a powerful influence for these districts in resulting tasks. Likewise note that a quickened calculation can be accomplished by embracing necessary pictures so as to diminish the duplication activities to those including just expansion as well as subtractions techniques. The methodology was applied into two main phases, the first to acquire face locale as well as furthermore mouth region was initiate from the lower-half of face, which are accepted in the mouth region of interest techniques.

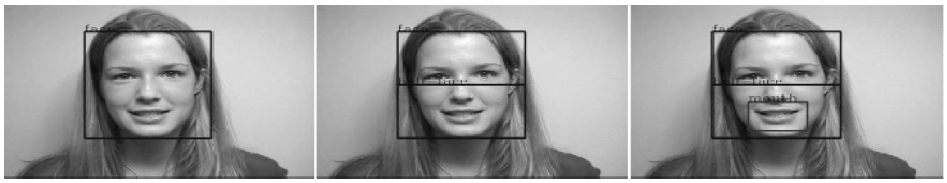

Figure 2. The face and mouth detection and extractions techniques such as (a) complete face-detection techniques (b) the half-face region of interest ROI extraction techniques and (c) the mouth region from lower part of the face respectively. Furthermore the figure depicted the face and mouth recognizable proof procedure and by expecting that are not solely is the estimation time expected to isolate from the actual extraction of the mouth district efficiently separated at this point this in like manner diminishes the peril of bogus area that can rise up out of the accidental course of action of an eye as a mouth.

2.2. Skin Detection Techniques

At the point when the mouth area have been isolated again with the wide determination of the computations are applied to the focus highlights. The gigantic number of elective highlights have been examined in composition of the skin detection techniques of the lip and non-lip areas have been separated with skin area and the purpose of clearing anyway several skins concealed of the pixels from of the photos as could be normal the situation being what it is in order to restrict the consideration on the remainder of the lip-hued region. Therefore a fitting concealing space ought to be looked over the wide combination of selections open together with red, green, maroon, dull, tone, inundation and worth. In this work have gotten the hue situation and value (HSV) concealing model for division model comes to closest replicating as well as individuals see skin concealing [10]. Therefore in the CUAVE pictures are taken in red, green and blue concealing space encompassing lights and surfaces headings comparative with light sources, therefore settling on HSV a decent decision for skin discovery techniques. The HSV shading channel is utilized to isolate the skin hued locale from the rest of the information picture. To discover suitable limit esteems to encourage partition, countless the pictures were inspected in the HSV shading space as well as segments ranges for the skin shading explored as appeared in the Figure 3 respectively. Therefore by using these limit esteems, a double picture is created in which those segments of the picture over the edge are made dark and the rest of colored are generated.

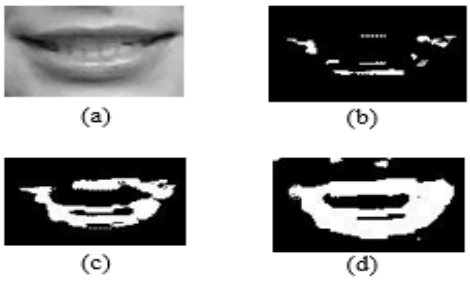

Figure 3: The skin-detection based HSV color filtration of binary images such as (a) original image (b) Feature Extraction techniques (c) Region of Interest Identification (d) Black and White colored segmentation.

2.3. Lip-Geometrical Feature Extractions Techniques

The fragmented picture containing the lip-district is then changed over to the double organization as well as shape extractions utilizing outskirt following are applied [11], and this produces an assortment of forms of the various-sizes and shapes disseminated along form of the twofold picture. Tests demonstrated that, of these forms, it is the biggest that normally most firmly coordinated the genuine lip layout. Despite the fact that this form intently pursues that of the lip, it stays a poor portrayal of the real state of the lip. This is on the grounds that the shape so created as well as straightforward polygon with the numerous non-converging edges as appeared in the above Figure 3. In any case, they are arched frame calculation can be pragmatic to acquire the curved polygon of the littlest zone. The figure 3(b) are explained the activity implemented by curved body calculation while the Figure 3(c) and Figure 3(d) explained the last polygon embodying every one of the focuses and such a raised polygon was utilized to speak the state of lip for consequent include extractions

3. LIP-GEOMETRICAL FEATURES TECHNIQUES IN MDTW

The lip-geometrical data from the lips shapes premise of lip-understanding procedure. However the performance can be economically upgraded utilizing data not from single picture and the grouping of highlight esteems acquired during the discourse expression. By finding the arched body form for every video outline, a period arrangement of highlight esteems can be inferred, as appeared in Figure 4. This data is put away as speaker models during preparing and for later use for lip perusing acknowledgment purposes in which the models are contrasted and recently created time arrangement.

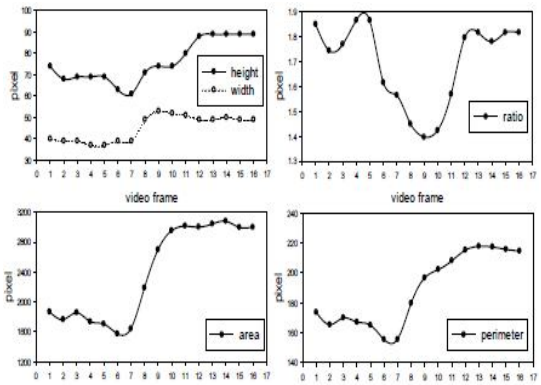

Figure 4: The dynamic lip in formation is obtained from1 by spoken by speakers of CUAVE database techniques

3.1. Dynamic Time Warping DTW Techniques in Audio Visual Speech Recognition

The DTW uses the dynamic programming procedure in order to enroll of common twisting among the betray game plan groups. Applying the configuration incorporates pairwise connection of component vectors as well as the improvement and weight of the zones are applied, and the outcome are perfect plan amongst the two component vectors progressions are delineated by misshaping limit procedures in audio visual speech recognition system and important its applications are shown in the present work of DTW divisions were resolved by using an Euclidean estimation techniques to generate the actual performance of DTW. Since five sorts of lip geometrical feature have been evacuated for the present lip getting structure and the DTW methodology just applies to single component course of action, a kind of decision blend is required to produces a strong relationship between the MDTW and audio visual speech recognition. This system prepared to relate all of the highlights at the same time, along these lines enhancing the synchronization between different features. To improve the lip perusing framework execution, examinations of order results accomplished utilizing mixes of lip geometrical highlights were completed with the correlation of Figure 5. The consequences of the work is appeared in Figure 5 and contrasted with the DTW of order results expand altogether for multi-included characterization when the performed utilizing MDTW, as this methodology fathoms of planning synchronization issues among highlights. As can be found in Figure 5 the mix of tallness, width, proportion, and Perimeter played out the best with 63% of order being effective utilizing MDTW yet the comparing the figure- being 52% for the DTW and the from the outcomes, it tends to be seen that basic proportions of stature as well as width and their proportion was adequate to speak to the lip-dynamic data as well as demonstrated appropriate for lip-understanding framework.

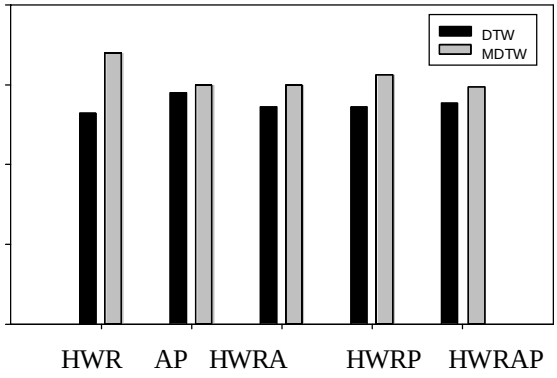

Figure 5: The word-accuracy by using the combinations of lip geometry features as well as classified as DTW and MDTW techniques.

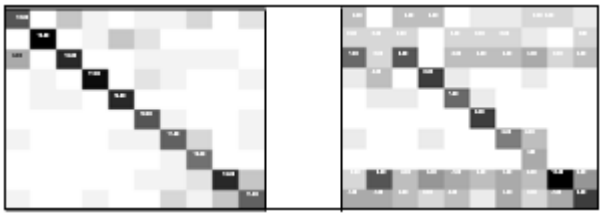

Figure 6: The confusion matrix for digits 0 to 9 by using the HWR features, (a) MDTW and (b) for optical flow.

4. THE HIDDEN MARKOV MODEL IN THE AVSR SYSTEM

HMMs are a well-known and fruitful way to deal with the factual displaying of perceptible discourse [12]. Since lip perusing additionally endeavors to perform discourse acknowledgment, yet in the visual-area, while the HMMs are frequently likewise utilized for visual-displaying. Therefore in this current work, we utilized in four unique sorts of the HMMs engineering dependent on the word as well as phoneme models. The three diverse word-models were executed within 3 states, 5 states and 7 states as appeared in the Figure 7 respectively. While phoneme model has different quantities of the states relying upon the substance to be named appeared in the Figure 8. Every Markov state is demonstrated utilizing exclusively Gaussian capacities with corner to corner covariance. The broadly utilized HTK library was utilized for the execution [13].

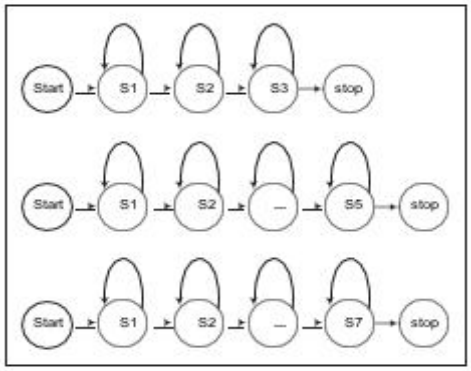

Figure 7: The 3-state top, 5-state middle and 7-state bottom-word recognitions models.

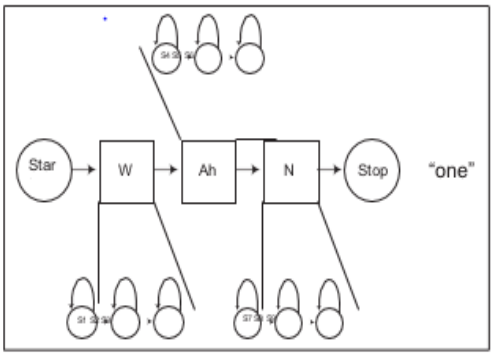

Figure 8: Complete phoneme recognition model technique.

Preparing was done utilizing just the first database video arrangements [14]. To explore order execution under conditions, where the head pivot as well as subject light-levels change, and various new groupings were made [15]. Towards the additionally show viability of novel approach outcomes were contrasted and an appearance-based method like those utilized in the numerous present down to earth acknowledge. In executed appearance-based methodology, each picture of mouth area to be located resized 64 x 64 as well as two-dimensional discrete cosign transform of the visual highlights were extricated from mouth locale, and to decrease computational exertion, just 16 discrete cosign transform of the low-recurrence constants were kept to the each picture. For each HMM state can be displayed utilizing 1, 2 and 3 blend Gaussian blend work. It is significant that a fitting Gaussian blend work is chosen for the specific applications, as this impact framework execution. The Figure 9 demonstrations presentation of geometrical based lip-perusing framework for the 3-states, 5-states, and 7-states HMMs, and the real indications the comparing consequences utilizing the appearance based DCT lip-understanding framework techniques.

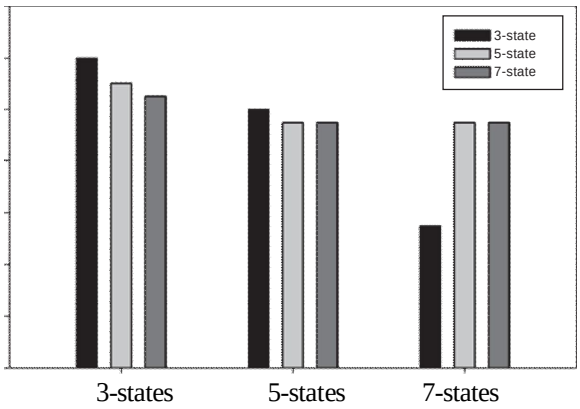

Figure 9: Lip-reading system using various HMMs architectures.

For both geometrical based as well as appearance based usage of outcomes demonstrated that presentation exacerbated as progressively by the Gaussian blend capacities were included and utilizing as solitary Gaussian capacity for the every Markov state was seen as most appropriate for lip-understanding framework. Therefore, the basic outcomes in accompanying segments were acquired for the HMMs designs that utilized a solitary Gaussian blend work. In genuine conditions, speakers will in general move their head-position while talking and as first endeavor to move-towards the performing lip-perusing in increasingly common circumstances. In this current work has examined the strength of new geometrical-based way to the head rotational developments. The Figure 10 shows the exhibitions of lip-perusing frameworks during head turn modifications.

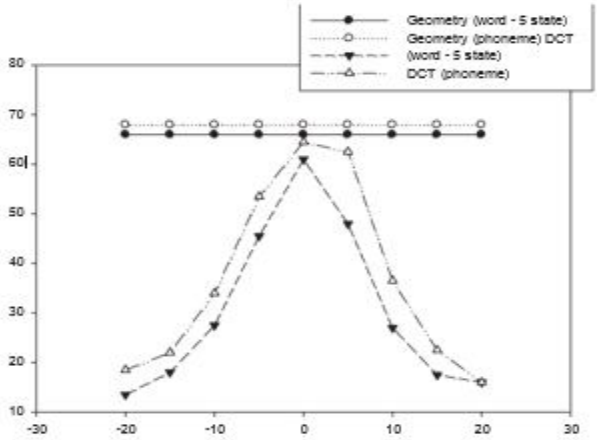

Figure 10: The geometric based and appearance-based lip-reading system through the head rotation changes.

5. CONCLUSION

This paper has proposed and assessed a lip-perusing framework utilizing of the dynamic data acquired from the lip-geometry methods. Examinations indicated of the novel MDTW strategy gives an elite lip include extraction procedure utilizing a skin shading channel an outskirt following technique and the count of the curved body systems. The outcomes have exhibited that the new lip division strategy acquainted is better capable with recognize the lip form than the dynamic shape system. Furthermore in this paper has likewise examined and presentation of a scope of lip-geometry features as well as various element blends. Another technique for fitting the activity of DTW to numerous highlights had the option to associate every one of the highlights all the while and in this way boost the synchronization between them. The commitment of the work introduced in this paper is twofold. Right off the bat, another technique is portrayed that can extricate lip geometry highlights dependent on the skin identification of HSV shading space and the estimation of the arched structure methodology. In the applications to perceive English digits starting from 0 to 9 as displayed in CUAVE database and

novel approach are performed to the superior to anything an appearance-based discrete cosign transform approaches. All characterization was done utilizing HMMs classifiers undoubtedly four sorts are design to examine and it was discovered that phoneme based model methodology. Assessment of the lip geometrical highlights indicated that stature and width data play out the best to speak to lip dynamic data. Notwithstanding, so as to accomplish great execution in genuine applications, examinations concerning the impacts that picture scaling and revolution has on the lip perusing framework should be led. In examination of both DTW and HMM with the cutting edge appearance-based techniques, in the novel geometrical based lip-perusing framework displays improved word acknowledgment exactness as well as starting tests that are shows in vigorous to pivot and splendor impacts, making it increasingly appropriate to be fused in multimodal discourse acknowledgment frameworks for use in uproarious situations.

REFERENCES

1. MI., Mohmand, A., Bhaumik, M., Humayun, Q., Shah (2019). An Assessment of the Visual Features Extractions for the Audio-Visual Speech Recognition. Published in International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), (Vol.8, No. 5), pp 2023-2028.
https://doi.org/10.30534/ijatcse/2019/27852019
2. Patterson, E.K., Gurbuz, S., Tufekci, Z., and Gowdy, J.N. (2002). CUAVE: A new audio-visual database for multimodal Human computer interface research. Proc. International Conference on Acoustics, Speech and Signal Processing, Orlando, FL, pp. 2017–2020.
3. MI., Mohmand, A., Bhaumik, M., Humayun, Q., Shah (2019). The Performance and Classifications of Audio-Visual Speech Recognition by Using the Dynamic Visual Features Extractions. Published in International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), (Vol.8, No. 5), pp 2049-2053.
https://doi.org/10.30534/ijatcse/2019/31852019
4. Rowley, H.A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1):23–38.
https://doi.org/10.1109/34.655647
5. Xu, Yan, Wang, Bin, Li, Jin Tao & Jing, Hongfang (2008) "An Extended Document Frequency Metric for Feature Selection in Text Categorization", "Information Retrieval Technology, Lecture Notes in Computer Science", Springer Berlin Heidelberg, Vol 4993, pp. 71-82, ISBN 978-3-540-68633-0.
6. Zhang, Y., Levinson, S., and Huang, T. (2000). Speaker independent audio-visual speech recognition. Proc. International Conference on Multimedia and Expo, New York, NY, pp. 1073–1076.
7. Zotkin, D.N., Duraiswami, R., and Davis, L.S. (2002). Joint audio-visual tracking using particle filters. In Press: EURASIP Journal on Applied Signal Processing.
8. H. Li and M. Greenspan (2011). “Model-based segmentation and recognition of dynamic gestures in continuous video streams,” Pattern Recognition., vol. 44, no. 8, pp. 1614–1628.
https://doi.org/10.1016/j.patcog.2010.12.014
9. L. Gillick and S. J. Cox, (2009). “Some statistical issues in the comparison of speech recognition algorithms,” in ICASSP 1989. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 532–535.
10. M. Gurban and J.-P. Thiran (2009). “Information theoretic feature extraction for audio-visual speech recognition,” IEEE Trans. Signal Process, vol. 57, no.12, pp. 4765–4776.
https://doi.org/10.1109/TSP.2009.2026513
11. S. W. Chin, K. P. Seng, and L.-M. Anger (2012). “Audio-Visual Speech Processing for Human Computer Interaction,” in Advances in Robotics and Virtual Reality, vol. 26, Springer Berlin Heidelberg, pp. 135–165.
12. N. Smolyanskiy, C. Huitema, L. Liang, and S. E. Anderson (2014). “Real-time 3D face tracking based on active appearance model constrained by depth data,” Image Vis. Computer., vol. 32, no. 11, pp. 860–869.
https://doi.org/10.1016/j.imavis.2014.08.005
13. S. H. Lee, M. K. Sohn, D. J. Kim, B. Kim, and H. Kim (2013). “Smart TV interaction system using face and hand gesture recognition,” in Digest of Technical Papers - IEEE International Conference on Consumer Electronics, 2013, pp. 173–174.
14. P. Dalka and A. Czyzewski, (2009). “Lip movement and gesture recognition for a multimodal human-computer interface,” in Proceedings of the International Multi-conference on Computer Science

and Information Technology, IMCSIT, vol. 4, pp. 451–455.

15. Bevilacqua, V., Ciccimarra, and A., Leone. G. (2008). Automatic Facial Feature Points Detection‖, Proceedings of the 4th international conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications - with Aspects of Artificial Intelligence, Shanghai, China, pp. 1142-1149
https://doi.org/10.1007/978-3-540-85984-0_137.