Computer Vision Architecture using Fusion Technique

Nidhi Srivastava
Amity University
Lucknow, India

Harsh Dev, Ph.D
PSIT College of Engineering
Lucknow, India

ABSTRACT

Humans want to communicate with the computers in the same way as they communicate with other humans. Speech is the most natural and spontaneous form of communication. Speech is bimodal in nature and it combines audio and visual information to enhance speech recognition rate especially under poor audio conditions. This paper proposes novel computer vision architecture using fusion technique. This architecture combines or fuses more than one modality using multi-agents. In this we have used two modalities- audio and video. The audio part extracts the speech of a person and the video part extracts the face and lip information of the person. Here, different agents process the modalities and the fusion agent fuses these modalities for effective and efficient automatic speech recognition.

Keywords

Computer vision; fusion; agent; architecture; multimodal; multi-agent

1. INTRODUCTION

Computers have become an indispensable part of our life. Not a single field exists today where computers are not being used. Machines have made the life of humans faster and easier. But, communication with computer is still not as easy as communicating with humans. Human to human communication is very natural and multimodal in nature.

For perceiving the world, people make use of vision, which is the primary sense of humans. The visual information plays an important part in interaction among people. The importance of vision is seen not only in interaction with humans only but also in interaction with machines. A machine that can see people is able to interact with humans in an informed manner. When we talk to a person along with the speech, the face, eyes, lip and hand movements also convey the information. Thus vision is important. Visual information has also been shown to be useful for improving the accuracy of speech recognition in both humans and machines. In computer vision, computer software and hardware are used to model and replicate the human vision.

In this paper we have given architecture for building a multimodal computer vision system using multi-agents. The agent is a new concept and technology in the field of software development. Agent has a characteristic that it can replace human and show the same intelligent behaviour as humans show [1]. Agents are small software entities which can be equipped with reasoning, learning and communication skills and display goal-oriented behavior [2]. Since an agent can take its own action, it helps in reducing the burden of a human being. Also, an agent if desired can become the advisor which can be very helpful. An agent assists the user by giving choices to them which can lighten their task. In case of building of a complex multimodal interfaces, a multi-agent architecture is considered appropriate. In multi-agent system, various agents interact with each other. This helps them in either achieving their individual objectives or else to manage the dependencies that result between them since they are situated in a common environment. A multi-agent system inherently supports parallel processing and may also provide means for distributed computing. A multi-agent system however may be quite complex as different agents work together in this [3, 4].

2. FUSION

In multimodal interactive systems, multimodal fusion is a critical and crucial step in combining and interpreting the various input modalities. It is a very delicate task.

Being a complex process, fusion needs to consider temporal issues as well as constraints presented by the use of different modalities. The nature of the input modalities used and the format of data supplied by agents connected to these sources of input play an important part in the fusion of the modalities. The fusing process depends on these variables [5, 6].

Integration of input signals from different modalities is little difficult and not an easy task. Each of the input modalities could carry partial yet complementary information that could be useful for the system. The problem with fusion is how and when to combine these inputs to decide what the user actually intends to communicate.

Different methods have been identified for fusion of the input modalities. Researchers have listed three methods: Data Fusion, Feature Fusion and Decision Fusion.

The lowest level of fusion is the Data fusion. Data fusion is considered where observations belong to the same type. It considers only the integration of raw observations. Raw data from different sources are combined to produce new raw data. Now, this data is expected to be more informative and synthetic than the inputs.

Feature-level fusion is a common type of fusion when tightly-coupled or strongly time synchronized modalities are to be fused. It assumes that each stream of sensory data is first analyzed for features, after which the features themselves are fused. The feature fusion strategy is generally preferred for closely coupled and synchronized modalities.

Decision-level fusion is based on the fusion of individual mode decisions or interpretations. In this fusion of loosely coupled modalities is carried out for example, pen and speech interaction. The modalities taken in decision fusion differ in the time scale characteristics of their features. In this, timing plays an important role and hence all fragments of the modalities involved are time stamped and further integrated in conformity with some temporal neighborhood condition [7,8].

3. PROPOSED ARCHITECTURE FOR FUSION

Finding a best possible fusion technique for a given combination of modalities is a tough task. Based on how these

modalities behave in a natural environment, a presumption can be made on the interaction and synchronization of these modes. However, it still remains necessary to explore multiple levels of fusion in order to determine the optimal combination of the desired modalities. This paper in detail explains the computer vision architecture for the agent framework given by us in [9]. There is a need to fuse the input modalities together for a successful recognition of speech. Integrating the information provided by different input modalities to get common information is a challenge for multimodal input fusion.

Diagrammatically, the architecture is shown in figure 1.

In this architecture, different input modality that is speech, face and lips, have been fused together.

3.1 Speech Feature Extraction

Speech has the potential to provide a natural user interaction model. However, it is not considered a very good interface due to prevalent disadvantages. Some of the disadvantages are ambiguity in spoken language, the limitations of current speech technology, etc. Speech content and voice can be interpreted as two different, though correlated, information existing in audio signals. The subject and recording condition determines the emotional information present in the voice [10]. The audio data is obtained with the help of mike. We already have an audio database where the different words are stored for recognition. Through mike the speech of the person is identified [20].

3.2 Visual Feature Extraction

Web camera helps in taking the image of the person standing or sitting in front and then detecting the face of the person. Face detection is a challenging task since there are many conditions that may vary. Each person has a unique face, meaning that each face looks different. The age of the person, eyeglasses, beard, moustache and make-up etc. all make a difference in the recognition of the face since all these factors have an effect on the face of a person [11, 12]. Thus detection of face is difficult. Many algorithms exist for face detection. We have proposed an algorithm [13] which detects the face of a person. Along with the detection of the face, the lip of the person is also identified. This is important due to the Mc Gurk effect in which if only speech is considered, the words may be misinterpreted. But, if along with speech, the face and lip movement of the person is considered then there is no confusion and the commands are easily and perfectly identified.

3.3 Agent assisted Context-Management & Multimodal Integration

The data obtained from speech and visual recognition not necessarily can be used as such. This data may need to be modified a little bit. So, with the help of agents we re-align and synchronize this data so that it can be used further. Now, the information from this Context Management & Multimodal Integration is passed over to the Inference fusion agent engine.

3.4 Inference Fusion Agent Engine

This Inference agent is an extension of the unimodal case as it accepts input from two streams rather than one stream. This inference agent seamlessly integrates the interpretation provided by the speech, face and lip. As we know, an agent is an autonomous entity. The environment provides information to the agent, which in turn processes it to reach a decision about further actions. Intelligent software agents have become popular also for extending the human interface beyond traditional direct manipulation interfaces. Software agent especially adapt to human computer interaction, because, user can exchange information with computer by using audio, video and lip motion used for various perception methods and must possess some cognitive function. Agents often provide services that are personalized to an individual's needs and desires, and they are often more appreciated when they provide assistance without being asked for. The goal of Inference fusion engine is to extract meaning from a set of input modalities and pass it to a user interface manager.

3.5 User Interface Manager & User

This user interface manager, with the help of inference agent, interprets the command and gives option to the user for selecting the right alternative. The user interface manager and user directly interact with each other. The user selects the correct option it wants and the command is executed. One of the options which is always given to the user is of discarding all the options. In this case the user can click on None of the above option. Once this option is selected, the message will be conveyed back to Inference fusion agent, which will re-assemble the two input modalities and will re-interpret the result and pass to the user interface manager. This process will be repeated three times, and at the end of the third trial if the user is still not satisfied, then the whole information is discarded and fresh input is taken from the user. There is a direct interaction between Inference fusion engine and User interface manager. The Inference fusion engine gives options to user which in turn either accepts or rejects the solution.

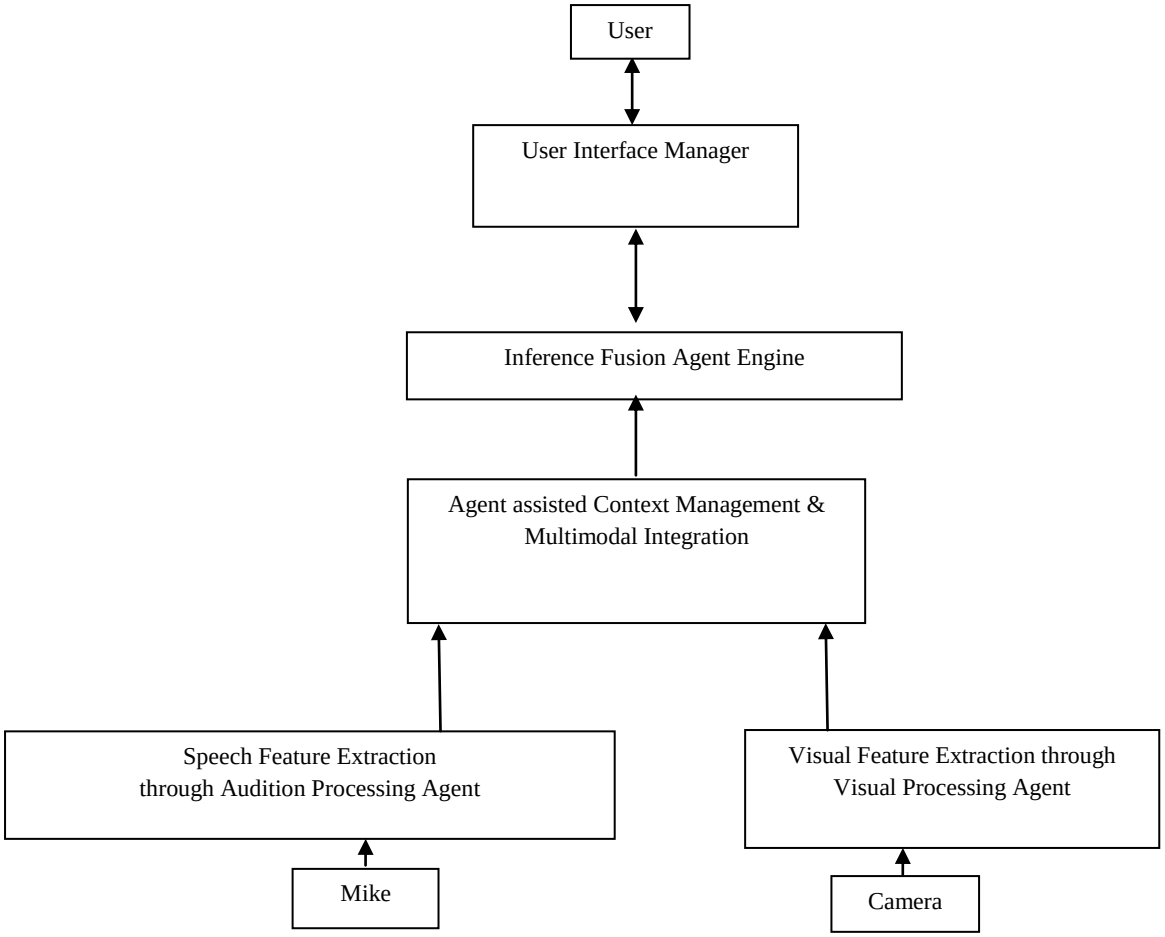

Fig. 1: Computer Vision Architecture

In both the cases, the final decision of the user is conveyed to the Inference fusion engine, which acts accordingly.

Many researchers have given different methods and have implemented them for fusion of the modalities. They have used Hidden Markov Model, Coupled Hidden Markov Model, Neural Networks etc. for fusion of the speech and visual channel. Various experiments conducted in [14, 15, 16, 17, 18, 19] show that audio-visual recognition systems have been able to outperform the audio-only recognition systems and visual-only recognition systems. Also, the audio-visual speech recognition scheme has much more improved recognition accuracy compared to audio-only and video-only recognition, especially in presence of noise. Thus, it can be said that this computer vision approach will make speech recognition more reliable and efficient.

4. CONCLUSION

This paper addresses novel computer vision architecture for fusion of the different input modalities for a multimodal system. This is a multi-agent system in which different agents have been used. All the agents work together and help in making the system an effective one. The audio and video modalities are processed by the audio and visual feature extraction agents respectively. Further, the data is re-aligned and synchronized by the agent-assisted and context-management agents. Inference fusion engine extracts meaning from a set of input modalities and fuses the two modalities-audio and video together and pass the information to a user interface manager. The user interface agent then interacts with the user and conveys the information. These software agents are very useful and help in multimodal human computer interaction. Thus, it can be said that this computer vision approach using multi-agents will make speech recognition more reliable and efficient.

5. REFERENCES

[1] Maximilian Kruger, Achim Schafer, Andreas Tewes, Rolf P. Wurtz, "Communicating Agents Architecture with Applications in Multimodal Human Computer Interaction", GI Jahrestagung, pp. 641-645, 2004.

[2] Elfriede I. Krauth, Jos van Hillegersberg, Steef L. van de Velde, "Agent-based Human-computer-interaction for Real-time Monitoring Systems in the Trucking Industry**"** IEEE Proceedings of the 40th Hawaii International Conference on System Sciences pp. 1-7, 2007.

[3] Cecilia Inks Sosa Arias, Beatriz Mascia Daltrini, "A Multi-Agent Environment for User Interface Design", Proceedings of the 22nd EUROMICRO Conference IEEE, Pague, pp. 242-247, 2nd -5th September 1996.

[4] J. Coutaz., "Interfaces homme-machine: un regard critique", Technique et Science Informatiques, 153-64, 1991.

[5] Simon C. Lynch (University of Teesside, UK) and Keerthi Rajendran (University of Teesside, UK)" A

multiagent approach to teaching complex systems development a hand book", 2011

[6] Shankar T. Shivappa, Bhaskar D. Rao, Mohan M. Trivedi "An Iterative Decoding Algorithm for Fusion of Multimodal Information", EURASIP Journal on Advances in Signal Processing, 2008.

[7] A. Corradini , M. Mehta, N.O. Bernsen , J.-C. Martin , S. Abrilian "Multimodal Input Fusion In Human-Computer Interaction on the Example of the NICE Project", Proceedings of the NATO ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management, NAREK Center of Yerevan University, Tsakhkadzor, Armenia, Kluwer, 18th -29th August, 2003

[8] Rajeev Sharma, Vladimir I. Pavlovic, Thomas S. Huang, "Toward Multimodal Human–Computer Interface", Proceedings of the IEEE, Vol. 86, No. 5, pp. 853-869, May, 1998.

[9] Prof. S.Qamar Abbas, Nidhi Srivastava, "Development of Framework for Automatic Speech Recognition", IJCSE, Vol. 4 No. 05 May 2012.

[10] Zhihong Zeng, Jilin Tu, Brian M. Pianfetti, Jr., and Thomas S. Huang, "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM", IEEE Transactions On Multimedia, Vol. 10, No. 4, pp. 570-577, June 2008.

[11] Erno Makien, "Face Analysis Techniques for Human-Computer Interaction", Tampere 2007.

[12] Nallaperumal K., Subban R., Krishnaveni.K, Fred L., Selvakumar K.R. Human Face Detection in Color Images Using Skin Color and Template Matching Models for Multimedia on the Web, IFIP International Conference on Wireless and Optical Communications Networks (IEEE), 7th August 2006.

[13] Nidhi Srivastava, Dr. Harsh Dev, Dr. Qamar Abbas, "Framework for Face Recognition", IJCA, Vol. 58, No.17, November 2012.

[14] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, Andrew W. Senior "Recent Advances in the Automatic Recognition of Audio-Visual Speech" Proceedings of the IEEE Vol. 91, No. 9, pp. 1306-1326, September 2003.

[15] Mustafa Nazmi Kaynak, Qi Zhi, Adrian David Cheok, Kuntal Sengupta, Ko Chi Chung, "Audio-Visual Modeling for Bimodal Speech Recognition" IEEE International Conference on Systems, Man, and Cybernetics, Tucson, AZ, Vol. 1, pp. 181-186, 2001.

[16] Trent W. Lewis, David M. W. Powers, "Audio-visual Speech Recognition using Red Exclusion and Neural Networks", Journal of Australian Computer Science Communications, Vol. 24 No. 1, pp. 149-156, January-February, 2002.

[17] Tieyan Fu, Xiao Xing Liu, Lu Hong Liang, Xiaobo Pi, Ara V. Nefian "Audio-Visual Speaker Identification Using Coupled Hidden Markov Models" Proceedings of International Conference on Image Processing, IEEE, Vol. 3, pp. III-29-32, 14th -17th September, 2003.

[18] Yashwanth H, Harish Mahendrakar and Sumam David "Automatic Speech Recognition using Audio Visual Cues" IEEE First Proceedings of the India Annual Conference, pp. 166-169, 20th – 22nd December, 2004.

[19] Jong-Seok Lee and Cheol Hoon Park "Robust Audio-Visual Speech Recognition Based on Late Integration", IEEE Transactions on Multimedia, Vol. 10, No. 5, pp. 767-779, August, 2008.

[20] Nidhi Srivastava, Dr. Harsh Dev, Dr. Qamar Abbas, "Speech recognition using MFCC and Neural Network", National Conference on Challenges & Opportunities for Technological Innovation in India, AIMT, February, 2013.