

LIPREADING BY OPTICAL FLOW CORRELATION

Glenn A. Martin

Computer Science Department
University of Central Florida
Orlando, FL 32816

Faculty Advisor: Mubarak Shah

INTRODUCTION

Voice recognition research has been around for years, and systems have been built with much success. However, in the presence of background noise, even the best of these systems will fail. Therefore, ways to prevent this using visual lipreading have been explored. Voice recognition supplemented with lipreading improves the accuracy of the recognition system.

Hazard [3], and Jeffers and Barley [4, 5] have studied lipreading for many years. They found that when people hard of hearing try to understand what somebody is speaking, they not only use what little they can hear, but also lip movements, facial expressions, and hand gestures, etc. This led to lipreading research in the computer vision field. If a lipreading system that performs well can be developed, it can combine with voice recognition systems to form a better overall recognition system.

We present a method for lipreading which uses sequences of dense optical flow. Optical flow is a vector field describing the image velocity induced due to the motion of an object, observer, or both. Because a dense optical flow field is used, data redundancy allows for better performance. To match model optical flow frames with input optical flow frames, they must be spatially warped, temporally warped, and correlated. Spatial warping locates the model within the input optical flow frames. Temporal warping finds a way to compare two optical flow sequences (especially if each is of different length). Correlation matching compares each model optical flow with the input optical flow using normalized correlation. Our recognition process uses spatial and temporally warping to perform correlation matching of the optical flow.

PREVIOUS WORK

Past methods of lipreading [7, 8] have used the shape of the mouth in solving word recognition, but clearly each person has a different mouth with respect to size and shape. Hence, the system must be trained for each speaker and will not work in a general environment. Other methods [2, 6] use the motion of the mouth instead since most sounds are created by different (yet unique) muscle movements. Thus, each person makes the same sound in basically the same way. However, these two methods use only a few selected points.

OUR APPROACH

This work uses lip movement for its basis of recognition, but the complete optical flow field is used in the comparison. Therefore, the redundancy of the data helps avoid poor and incorrect matches. Optical flow is found for each sequence by Anandan's algorithm [1] because of its accuracy. Then, spatial and temporal warping programs find the best way to compare sequences. The spatial warping determines where the lips are located in an image since it is nearly impossible to have each person at the exact same spot for each sequence. The temporal warping is necessary since people talk at different speeds creating sequences of lip movements with a different number of frames for a specific sound. A correlation matching program finds the best match between the optical flow frames of the input and the optical flow frames of each model using this mapping.

Spatial Warping

The spatial warping program uses a simple matching routine based on correlation. A small window around the lips in *each* model optical flow frame is compared to *each* input optical flow frame in their respective sequences. This will build a two-dimensional array with the model flow frames along the x-axis and the input flow frames along the y-axis. Each element of the array will contain the best-match correlation value along with its corresponding (x,y) location. As one would expect, the comparisons among two frames that should never match up will be very poor.

Figure 1 shows how the two-dimensional array is built. This example has a model sequence of three optical flow frames, and an

Input	3	(0, 30) 15	(0, 27) 19	(0, 25) 89
	2	(0, 45) 27	(0, 27) 78	(0, 25) 16
	1	(0, 57) 96	(0, 27) 70	(0, 25) 45
		1	2	3
		Model		

Figure 1: Implementation of Spatial Warping in 2-D Array. For each optical flow frame (1–3) in the model sequence, its pre-defined window is located in each optical flow frame of the input sequence using correlation. The location and value of the best match (the maximum of the correlation values) is stored in this array.

input sequence of three optical flow frames. For each frame in the model sequence, its pre-defined window (of the lips) is located in each input frame. To accomplish this, the window is placed in all possible places in each input frame and correlated with the area it covers. The location and value of the best match (the maximum of the correlation values) are stored in the array.

Temporal Warping

The temporal warping program finds the best way to compare the model and input sequences. This algorithm is very critical to the whole matching process. It is needed in order to compare two sequences that might be shorter than each other. This happens with people who talk faster than others resulting in sequences with a different number of frames.

Presently, the temporal warping program is based on Sakoe and Chiba’s method [9]. To accomplish this, it takes the output array of the spatial warping program, and considers it as an adjacency matrix of a graph. It then tries to find the best path through this graph using a standard algorithm. The list of points that create this best path are then stored in a file for correlation matching.

Correlation Matching

The correlation matching program is the simplest. For the given input sequence and *each* model sequence, it sums the correlation

values found in the spatial warping program that are defined by the path determined by the temporal warping program, and finds the average. Since correlation is used, the model with the maximum average is the best match.

RESULTS

Images for each sequence were collected at a rate of 30 frames per second. Different people supplied the vocabulary and one was arbitrarily chosen as the model. During this process, fairly good lighting was used in a computer lab environment. Each resulting image was then cropped from 640x320 to 256x256 roughly centered around the lips to “zoom-in” without losing focus. Regardless of the others, each sequence was captured from the beginning of the word to the end.

To start, sequences for each model were taken, and the optical flow was found. As each input sequence is created, its optical flow is computed, and then spatially and temporally warped with *each* model. Finally, it is compared with each model to find the best match.

Figure 2 shows two images of a sequence and the corresponding optical flow between images. This is only one letter of three that were tested. Figure 3 shows how each of the three inputs (d, h, and t) matched with the three models (d, h, and t). Although only one input matched correctly (33%), this can be explained by visemes which are groups of letters that are visually similar. For example, the letters “d” and “t” appear almost exactly the same. Therefore, they are often grouped together and one matching with the other is considered acceptable. Therefore, our match rate increases to 66%.

CONCLUSIONS

The research described provides insights to the problem of lipreading. For instance, spatial warping may not be necessary if the location of the lips is already known. An error of a few pixels should not hurt the match. Temporal warping, however, is the most important function since recognition systems must compare sequences which have different numbers of frames. Finally, normalized correlation seems to generate the appropriate matches, but further testing will be required to obtain confirmation.

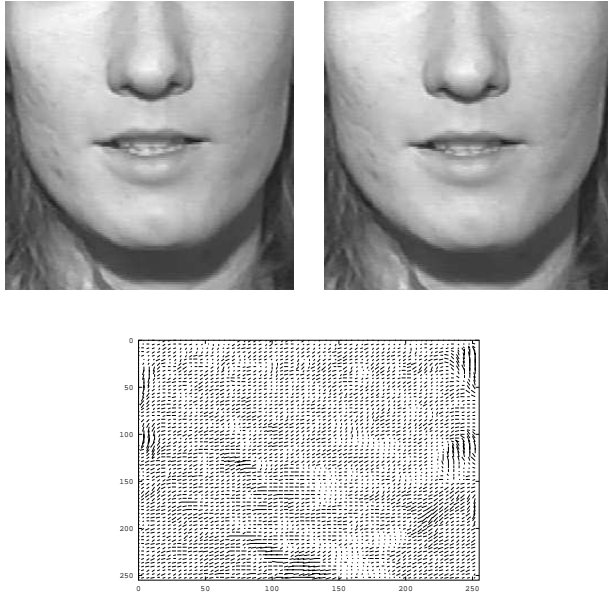


Figure 2: Two images from one sequence, and the corresponding optical flow between them.

Input	Match
d	d
h	t
t	d

Figure 3: Results of matching the three inputs (d, h, and t) with the three models (d, h, and t). Although only a 33% match rate is found, viseme grouping raises it to 66%.

References

- [1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, January 1989.
- [2] Kathleen E. Finn and Allen A. Montgomery. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, October 1988.
- [3] Elizabeth Hazard. *Lipreading*. Charles C Thomas, 1971.
- [4] Janet Jeffers and Margaret Barley. *Speechreading (Lipreading)*. Charles C Thomas, 1971.
- [5] Janet Jeffers and Margaret Barley. *Look, Now Hear This*. Charles C Thomas, 1979.
- [6] Kenji Mase and Alex Pentland. Lip reading: Automatic visual recognition of spoken words. *Optical Society of America Topical Meeting on Machine Vision*, 14:1165–1570, June 1989.
- [7] Eric Petajan, Bradford Bischoff, David Bodoff, and N. Michael Brooke. An improved automatic lipreading system to enhance speech recognition. In *SIGCHI '88: Human Factors in Computing Systems*, pages 19–25. ACM Press, October 1988.
- [8] Eric D. Petajan. Automatic lipreading to enhance speech recognition. In *Computer Vision and Pattern Recognition*, pages 40–47. IEEE Computer Society Press, June 1985.
- [9] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-26(1):43–49, February 1978.

References

REFERENCES

- [1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, January 1989.
- [2] Kathleen E. Finn and Allen A. Montgomery. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, October 1988.
- [3] Elizabeth Hazard. *Lipreading*. Charles C Thomas, 1971.
- [4] Janet Jeffers and Margaret Barley. *Speechreading (Lipreading)*. Charles C Thomas, 1971.
- [5] Janet Jeffers and Margaret Barley. *Look, Now Hear This*. Charles C Thomas, 1979.
- [6] Kenji Mase and Alex Pentland. Lip reading: Automatic visual recognition of spoken words. *Optical Society of America Topical Meeting on Machine Vision*, 14:1165–1570, June 1989.
- [7] Eric Petajan, Bradford Bischoff, David Bodoff, and N. Michael Brooke. An improved automatic lipreading system to enhance speech recognition. In *SIGCHI '88: Human Factors in Computing Systems*, pages 19–25. ACM Press, October 1988.
- [8] Eric D. Petajan. Automatic lipreading to enhance speech recognition. In *Computer Vision and Pattern Recognition*, pages 40–47. IEEE Computer Society Press, June 1985.
- [9] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-26(1):43–49, February 1978.