# REAL-TIME SPEAKER IDENTIFICATION FOR VIDEO CONFERENCING

S.Saravi, I.Zafar, E.A.Edirisinghe, *R.S.Kalawsky
Digital Imaging Research Group, Department of Computer Science
*AVRRC, Department of Electrical & Electronic Engineering
Loughborough University, Loughborough, LE11 3TU
s.saravi@lboro.ac.uk, i.zafar@lboro.ac.uk, e.a.edirisinghe@lboro.ac.uk, r.s.kalawsky@lboro.ac.uk

## ABSTRACT

Automatic speaker identification in a videoconferencing environment will allow conference attendees to focus their attention on the conference rather than having to be engaged manually in identifying which channel is active and who may be the speaker within that channel. In this work we present a real-time, audio-coupled video based approach to address this problem, but focus more on the video analysis side. The system is driven by the need for detecting a talking human via the use of computer vision algorithms. The initial stage consists of a face detector which is subsequently followed by a lip-localization algorithm that segments the lip region. A novel approach for lip movement detection based on *image registration* and using the *Coherent Point Drift (CPD)* algorithm is proposed. Coherent Point Drift (CPD) is a technique for rigid and non-rigid registration of point sets. We provide experimental results to analyse the performance of the algorithm when used in monitoring real life videoconferencing data.

**Keywords:** Speaker Identification, Coherent Point Drift, Lip Movement Detection

## 1. INTRODUCTION

One of the most serious problems in video conferencing facilities available today is the difficulty in determining who is speaking amongst a large number of conference participants. There is a strong need for providing remote users with a close-up of the current speaker which automatically tracks when a new participant begins to speak. A number of attempts have been taken in literature to address the above issue.

In [3] a speaker identification system was presented that was supported by dynamic visual data from video sequences including the lip region. The geometrical features of the lips analysed included data about the shape and intensity information of the lips. This person identification system based on Hidden Markov Model (HMM) was proved to reach accuracy figures of up to 72.2% [3]. Some researchers used statistical visual data in recognition systems that use morphological filtering for a facial/eye localization and subsequently a matching algorithm for identity confirmation [4]. In [5] authors used Principal Component Analysis (PCA) for facial feature extraction for identity confirmation and in [6] authors used visual information from different components in the face for speaker identification. In [7] authors developed a speaker verification system based on dynamic lip contour features using Linear Discriminant Analysis (LDA) in combination with principal component analysis. This study was developed to combine audio-visual data by late integration using mixed densities of Gaussians [8]. [9] presented a system using multimodal visual information from a video sequence. The modalities, face, voice and lip movement, were merged using voting and opinion synthesis. Another related research has exploited dynamic visual information for speaker recognition [10], and used multimodal information for speaker identification [11]. In [12] the authors proposed a technique based on HMMs to put together multimodal data considering synchronization and weights for different modalities. In [13] a dynamic video-only biometric system was implemented with a robust and automatic method for tracking speakers' face.

Despite all the above efforts speaker identification still remains an open research problem due to issues related to speed and accuracy of the existing systems. Real-time techniques capable of performing well under different levels of facial illumination, shadows etc, especially on the lip region which is very closely analysed for possible detection of movement, is required. Our approach summarized below attempts to extend the state-of-art in speaker detection by proposing a robust, real-time system.

Real-Time Image and Video Processing 2010, edited by Nasser Kehtarnavaz, Matthias F. Carlsohn, Proc. of SPIE Vol. 7724, 77240D · © 2010 SPIE
CCC code: 0277-786X/10/$18 · doi: 10.1117/12.854846

In a typical distributed video conferencing environment multiple microphones and cameras located in multiple meeting rooms capture voice signals and video streams. The audio signals arriving from the microphones can be easily analyzed and classified as human voice or noise using a human voice detection approach. The presence of human voice in a channel verifies that the associated video stream would have captured speaker. Then the selected video stream is analyzed with a face detection algorithm [1]. This algorithm locates faces in each frame of the video stream using Haar like features. The face detection stage is followed by lip localization. We assume based on human facial anthropometric data that the human mouth is located in the bottom third of a human face. Therefore a simple subdivision of the facial area can lead to lip localization. The last step is to detect the lip movements, which is a challenging task. The basic idea of the lip movement detection algorithm is to compare corresponding points between two views of a lip and measure the relative movements and subsequently classify as to whether the shape of the lip has changed significantly, i.e. talking. In the proposed work we use **Coherent Point Drift (CPD)** [2] for the above purpose. For each comparison, variance is calculated between the locations of the corresponding points. As the average variance for 'not talking' frames is considerably lower comparing to the 'talking frames' a robust classification can be obtained. Due to the robustness of the CPD algorithm to noise and limitation of edge detection algorithms, we prove that the proposed approach to lip movement detection is robust, whilst performing real-time.

Section 2 explains the proposed system which includes detailed discussions on face detection, lip localization and lip movement detection approaches utilized in our design/implementation. Section 3 provides experimental results and an analysis. Section 4 concludes with suggestions for further improvement.

## 2. PROPOSED SYSTEM

Figure 1 illustrates the block diagram of the complete speaker detection system which includes the audio processing blocks which are not discussed in detail in this paper. The speaker identification system starts with analyzing audio signals which come from different channels. The signals are checked and categorized as human voice and noise using a pre-trained algorithm that is capable of identifying human voice from other audio. As a result of this audio processing the video stream through which the human voice comes from is selected. The selected video is analyzed using three key stages of computer vision algorithms as follows:

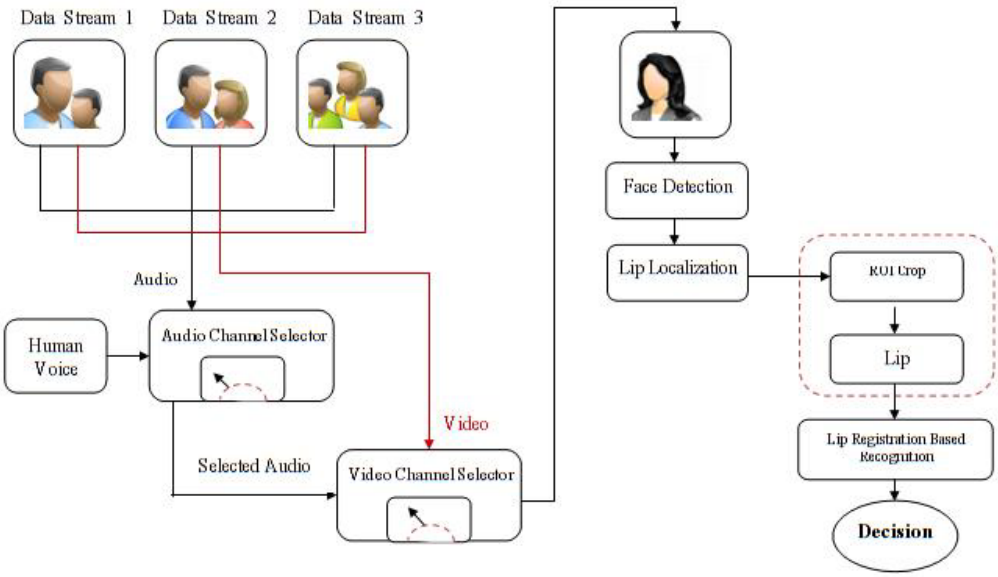



Figure 1. Speaker Identification System

## 2.1 FACE DETECTION

Face detection is the first step in automatic face recognition. For this purpose the OpenCV face detector was used [1]. In this solution the aim is to build a classifier cascade with generally high correct detection and low false detection rate, although every stage in the cascade has a strong classifier including a group of weak classifiers. All weak classifiers are bounded to a unique Haar-like feature. To select the specific Haar features to use, and to set threshold levels, Viola and Jones used AdaBoost. AdaBoost procedure is used to choose better features to form a stronger classifier. It merges a number of weak- classifiers to create one strong classifier. Weak classifier is the classifier which gets the correct answer more often than random guessing. Each of the weak classifiers adds more correction to final answer. A classifier is trained with a number of different views of a face, which are balanced to a unique size and negative images of the same size. Two sets of grayscale sample images are used in the training procedure. One set includes human face samples and the other set includes random non-face samples.

After training, the classifier can be applied to a region of interest in an input video or image. The classifier results will return "1" if the detected region is the face; otherwise it returns "0". The search is based on the "window search" which moves across the whole image or each frame of video and checks each and every location using the trained classifier. Viola and Jones named this filtering sequence a cascade classifier. The proposed algorithm in OpenCV could not satisfy the results needed. There were some frames with no or false detection, therefore codes were modified to get the desired results. The first step was to convert the video format to 4:2:0 YCbCr (YV12). Next step was to feed the program with video samples. The faces in the video were enclosed with a red rectangle.

During the execution some non face or false areas were detected. As a solution, it was assumed that always the biggest rectangle detected is the correct answer. The algorithms were extended to solve the problem. In each frame two points are defined for identifying rectangles, x and y of the top left and x and y of the bottom right corners of the enclosing rectangle. Using these values the area of the rectangle are calculated and the values which generated the biggest area are written to a text file as output. At this stage only one rectangle in each frame is shown, but now the question is what if the current rectangle is not the face. The idea is that in a video which frame rate is 30 per second, the location of the face will not be moving more than about 20 pixels (in any direction) due to practical reasons, in consecutive frames. Thus the rectangular region in current frame will be in neighboring location to the rectangular region of the previous frame. Reading the video and text file, the position of rectangle points (i.e., x and y values of top left and bottom right points of rectangle) from current frame is compared with the points from previous frame. If the difference is bigger than 20 pixels, the point values from current frame will be replaced with the point values from previous frame. The overall idea is to remove outliers of the facial locations within a video based on the assumption that the movements are smooth.

## 2.2 LIP LOCALIZATION

Next step is lip localization. Lip movement detection initialization involves an approximate estimation of the lip location. The lip locator has to initially find the location of the mouth in human face in a digital image or a video frame.

Most lip localization techniques are feature or edge based. Disadvantage of feature based approach is the low speed. Numerous features must be tested on several parts before a correct location of the mouth can be found. Thus these techniques will not be suitable for implementing real time systems when added to the complexities of the other important stages of a speaker detection system. Lips can also be found using edge detection. The reason of detecting sharp changes in image brightness is to notice important changes in properties of the image. The outcome of applying an edge detector to an image will be a set of linked lines that specify the borders of objects. Hence, applying an edge detector to an image considerably decreases the quantity of information to be processed and as a result eliminates the data that in less important, while keeping the significant information of the image. Edges detected in images can often lead to poor results due to the edge lines being not connected, i.e. fragmented, and also due to false detections or missed edges. Thus edge detection based approaches for lip localization is not robust.

Due to the above reasons for lip localization we adopt a simple approach based on human facial anthropometric ratios. Once the face detector detects the facial region it is subdivided horizontally into three regions and the lips are assumed to be located in the bottom third. In a similar way, the face is divided by four vertical lines the lips are located horizontally to be within the two middle quarters of the face (Figure 2). Based on our experiments this approach was both simple enough to be real-time in operation and robust to poor lighting conditions etc. Once the lip area is localized more detailed analysis of the smaller space can be used to detect possible lip movement. This will not be computationally extensive as the area being processed is substantially smaller as compared to the facial area.

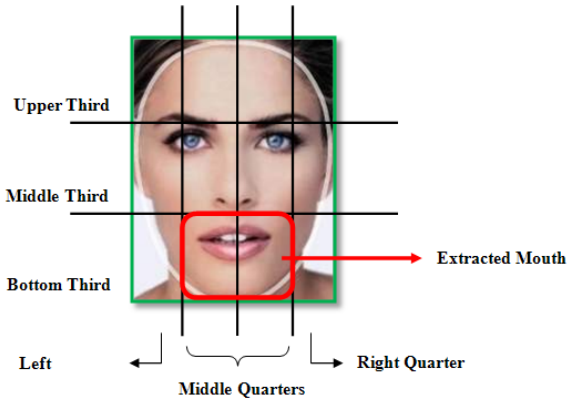



Figure 2. Facial shape and division

Using the above approach, the lip location can be extracted from the detected face. As a result the lip area will be localized and can be used in next step, lip movement detection.

### 2.3 LIP MOVEMENT DETECTION

This stage enables the identification of a talking person by comparing of the shape of the lips to a closed mouth or via temporal analysis of shape changes that occur in a lip within a sequence of frames. The approach proposed is based on the idea of CPD [2].

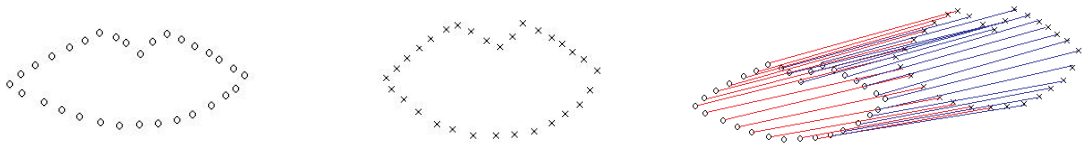
CPD is based on 'Registration of point sets', which is has applications in areas such as robot navigation; image guided surgery, motion tracking, and face recognition. The aim of point set registration is to create connections among two sets of points and to find the modification that links one point set to the other. Indeed, it is the base part in tasks such as object positioning, stereo matching, point set connection, image segmentation and shape or pattern matching. The problem of image registering is finding significant connections among two point sets and to evaluate the principal transformation that links one point set to the other. Point set includes the points which are features, which mostly are the positions of significant points taken out from an image. Every geometrical feature can be corresponded to a point set. There are two methods for registering an image, rigid or non-rigid, based on the transformation model principal. The main attribute of a rigid transformation is "distances preserved". The basic non rigid transformation is affine, which also enables scaling and skews. Effective algorithms exist for rigid and affine registration [14] and [15].

For image registering a probabilistic method has been introduced in [2] for point set registration that is named as the Coherent Point Drift (CPD) method. Given two point sets, a Gaussian Mixture Model is fitted to the first point set, whose Gaussian centroids are initialized from the points in the second set. For the process of adapting the Gaussian centroids from their initial positions to their final positions as a temporal motion process, and impose a motion coherence, a constraint over the velocity field has been created. The assumption is that points close to one another tend to move coherently. This constraint penalizes derivatives of all orders of the underlying velocity field (thin-plate spline only penalizes the second order derivative). A solution has been derived for the velocity field through a variation approach by maximizing the likelihood of GMM penalized by motion coherence. Further an EM algorithm for the penalized ML optimization with deterministic annealing has been derived. This method is a true probabilistic approach and is shown to be accurate and robust in the presence of outliers and missing points, and is effective for estimation of complex non-linear non-rigid transformations.

In simple words, Coherent Point Drift (CPD) is a new technique for non-rigid registration of point sets. The whole idea is to move one point set coherently to align with the second point set (Figure 3, (a) and (b)). The CPD method finds both the non-rigid transformation and the correspondence distance between two point sets at the same time without having any previous declaration of the transformation model except motion coherence (Figure 3, (c)).

The CPD idea is utilized in the proposed work for lip motion detection. The idea is to first carry out an edge detection (e.g. Canny edge detector) on the localized lips of all frames and to compare the edge points of all lips in a set of frames with the edge point set of a reference non talking lip and check the amount of change between their correspondences based on the CPD approach (see Figure 3). If the variance in the deviations is low, it means that the lip has not moved or

the movement cannot be considered as sufficiently large to classify the lip as a talking lip and if the result is high, it means that the lip shape has been changed signifying a talking lip. For achieving this idea, the first frame of video (which is assumed to be a non talking lip) is chose as the reference image. Then using the canny edge detection the edges of the lip will be taken out and saved as the point sets. For the images which will be registered to the reference image the same action will be performed to obtain the point sets. All upcoming frames will be compared to the base image.



(a) (b) (c)

Figure 3. (a) Point Sets of Non-Talking Mouth, (b) Point Sets of Talking Mouth, (c) Correspondence Distance between Point Sets

Each point from the second image will be moved to the base point and this movement will result in a set of correspondence distances. Then, the variance of correspondence distance is calculated for all frames. Plotting and comparing the variance of whole video, it can be easily seen that there is a significant difference between the results of talking lips and non talking lips. Obviously if the person is not talking the lips will not move significantly. Thus the correspondence distance will be small and vice versa. By calculating the average of all variance which resulted from annotated training videos it was found that average a variance values below 15 signifies a non-talking lip where as above this threshold it can be classified as a talking lip.

## 3. EXPERIMENTAL RESULTS AND ANALYSIS

A total of ten video samples were used in out testing. These samples depicted conversational videos obtained under various illuminations, recording people with different skin colour and tone, still and moving faces, talking faces and non talking faces or a combination of the above mentioned situations.

### 3.1 FACE DETECTION

Different videos as described above were used to test the OpenCV face detector. As mentioned previously this approachresulted in some faulty detections (see Figure 4).

(a) (b) (c)

Figure 4. (a) Faulty Face Detection, (b) Only One Rectangle Detected Per Frame, (c) Modified Rectangle

As it can be seen in the sample result illustrated in Figure 4, more than one rectangle was identified in a given frame. The first step was to eliminate all the rectangles but one which is considered to be the most appropriate. For this reason, the area of all rectangles detected in each frame was calculated using the x and y values of the top left point and bottom right point. The areas of the rectangles were compared with the areas of the rectangles detected in the previous frame, i.e.

temporarily and the rectangle with minimum deviation of size and location was assumed to be belonging to the detected face and the other was ignored. This algorithm allows multiple faces tp be detected and corrected in parallel. However in our experiments the implementation was limited for correcting the location of a single detected face. The correct rectangle was marked / coloured in green (see Figure 4).

### 3.2 LIP LOCALIZATION

In order to localize the lips, the detected face was cropped by separating out the bottom third of face (horizontally) and middle two quarters (vertically). Figure 5 indicates the extracted lip part from the detected face.

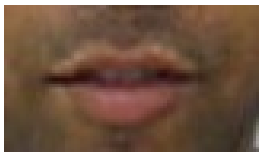


Figure 5. Successfully Cropped Lip Area

### 3.3 LIP MOVEMENT DETECTION

The results from lip localization were used in the lip movement detection stage. The first frame which was assumed to be a non talking lip was chosen as the base image. Then using the Canny Edge Detector, the boundaries of the lips were found (Figure 6). The same process was applied for all frames. The edges are represented by a set of pixels; therefore, they can be used as point sets which can subsequently be analysed using the CPD based approach.

(a) (b)

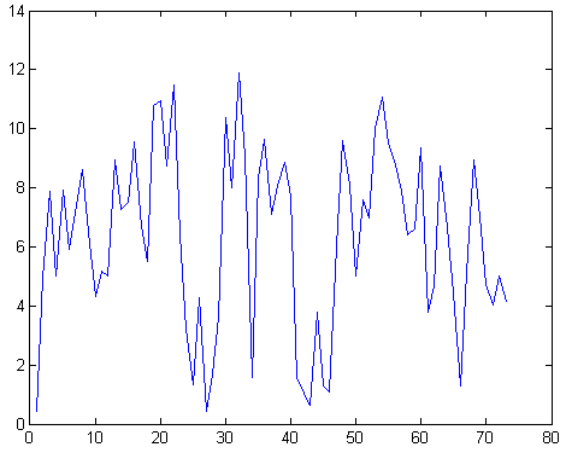
Figure 6. (a) Sample of Cropped Lip Area, (b) Detected Lip Boundaries

The challenging task in this section was edge detection. The canny Edge Detector did not perform as expected in some images. Setting the threshold was difficult as finding the best threshold (i.e., which gives the best detection of edges) to be used in each frame was difficult and would have resulted in significant overhead. Therefore a single threshold range was used for all the frames which at times resulted in some non-optimal edge detections such as broken edge representations and some edges not been picked up etc. However the CPD approach automatically detects the sets of corresponding pairs of points between the edge representation of the test and reference frames, ignoring edge points which do not find a corresponding edge point being detected in the other image. The edge point set of each detected lip in frames was compared to the edge point set of the reference image's lip and the correspondence distance between two point set was calculated. For each lip a set of correspondence distances are obtained. The variance of each point set was calculated and plotted. Observing this plot (see figure 7, 8 and 9 (b)) talking and no-talking lips can be identified by checking against an average variance threshold value of 15. Lips that resulted in average variances above 15 were classified as talking lips whereas below were considered as no talking lips.

The following tests were done on video samples taken individually from a talking and not talking person. The table in Figure 7 indicates the variance of correspondence distance in each frame of a non-talking video. The graph illustrates the variations clearly. It can be easily seen that the variance of each frame for the non-talking video is very low. It is noticeable that the mean of variances for a non-talking video is always below 15. The table in Figure 8 specifies the variance of correspondence distance of each frame of a talking video. The corresponding graph draws attention to the fact that high values (above 15) of average variances are depicted.

| Variance of Corresponding Distance for "Still" Video | | | | |
|---|---|---|---|---|
| 0.4369 | 9.5727 | 7.9871 | 1.0750 | 3.7762 |
| 4.9819 | 6.6978 | 11.9011 | 6.0395 | 4.6752 |
| 7.8971 | 5.5245 | 9.0711 | 9.5866 | 8.7358 |
| 5.0170 | 10.7884 | 1.5821 | 8.0947 | 6.6613 |
| 7.9375 | 10.9437 | 8.3186 | 4.9978 | 4.4416 |
| 5.9088 | 8.7256 | 9.6574 | 7.5840 | 1.2785 |
| 7.1988 | 11.4776 | 7.1016 | 6.9594 | 5.1536 |
| 8.6245 | 6.3971 | 8.0722 | 10.0020 | 8.9503 |
| 6.4699 | 3.2315 | 8.8776 | 11.0709 | 7.0293 |
| 4.3187 | 1.3426 | 7.6989 | 9.5097 | 4.7386 |
| 5.1851 | 4.2939 | 1.5729 | 8.7676 | 4.0468 |
| 5.0251 | 0.4130 | 1.0393 | 7.8786 | 5.0150 |
| 8.9503 | 1.6902 | 0.6533 | 6.3959 | 4.1491 |
| 7.2586 | 3.6085 | 3.8014 | 6.5985 | |
| 7.5284 | 10.3956 | 1.3067 | 9.3396 | |
| **Mean of Variance of Corresponding Distance: 6.2881** | | | | |

(a)

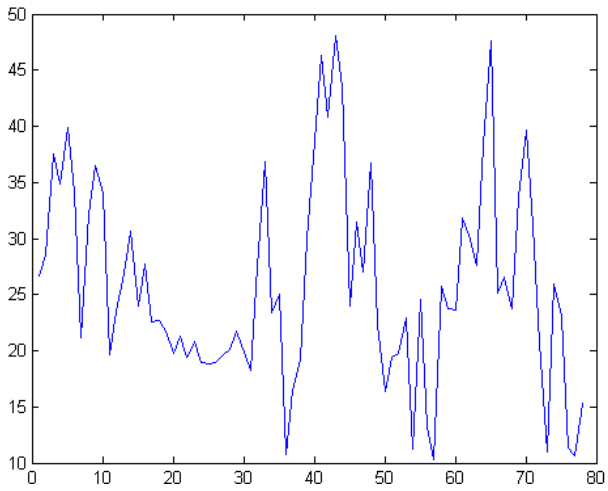



(b)

Figure 7. (a) (Left to Right) Variances of Correspondence Distances, (b) Plotted Variances of Correspondence Distances

| Variance of Corresponding Distance for "Talking" Video | | | | |
|---|---|---|---|---|
| 26.209 | 28.5329 | 37.5262 | 34.8882 | 39.8508 |
| 34.1631 | 21.1533 | 32.1238 | 36.4393 | 33.9797 |
| 19.6882 | 23.5028 | 26.8475 | 30.6281 | 24.0042 |
| 27.7730 | 22.5713 | 22.6922 | 21.7208 | 19.7109 |
| 21.3237 | 19.3798 | 20.8355 | 18.9148 | 18.7812 |
| 18.8941 | 19.5306 | 20.1097 | 21.7513 | 19.8901 |
| 18.1954 | 27.5239 | 36.8231 | 23.3371 | 25.0352 |
| 10.7692 | 16.5397 | 19.2144 | 29.9694 | 38.7377 |
| 46.3544 | 40.7686 | 48.0673 | 43.5273 | 23.9449 |
| 31.4527 | 27.0735 | 36.6982 | 22.3877 | 16.4230 |
| 19.4384 | 19.8182 | 22.8714 | 11.1962 | 24.5385 |
| 13.2058 | 10.2361 | 25.7002 | 23.7931 | 23.5954 |
| 31.7812 | 30.3110 | 27.5770 | 38.3885 | 47.6030 |
| 25.1865 | 26.5777 | 23.7654 | 33.7151 | 39.6288 |
| 31.0835 | 20.6688 | 11.0061 | 25.9198 | 22.9828 |
| 11.3495 | 10.6443 | 15.8840 | | |
| **Mean of Variance of Corresponding Distance: 25.8287** | | | | |

(a)



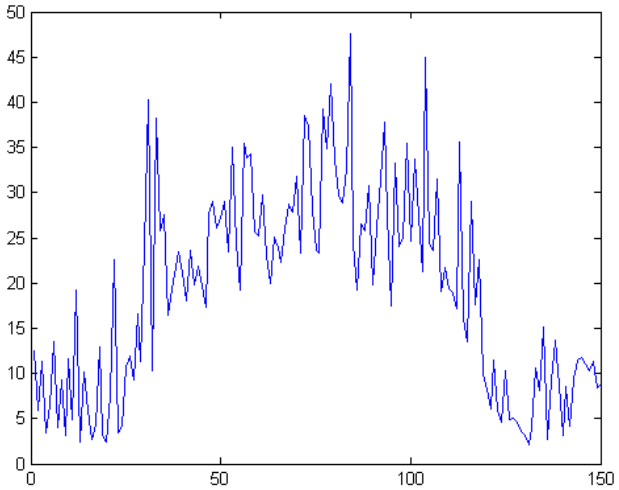

(b)

Figure 8. (a) (Left to Right) Variances of Correspondence Distances, (b) Plotted Variances of Correspondence Distances

The following results are from video which contains a person who first does not talk and then talk for a short while before stopping once again. The divisions of talking and non-talking frames were as follows: fames 1 to 30, non-talking, frames 31 to 114 talking and frames 115 to 150 not talking. Looking at the values of variances from the table and graph in figure 9 closely, it is seen that our thresholding approach allows the proper identification of the lip status change locations, accurately. Note that sudden unexpected changes in the variances are related to the incorrect edge detections or imperfectly cropped lip locations, which includes some points selected ftom the background. The averaging of the variances before comparing with a threshold resolves this complication.

| Variance of Corresponding Distance for "Non-Talking, Talking and Non- Talking" Video | | | | | |
|---|---|---|---|---|---|
| 12.5402 | 5.8915 | 11.2944 | 3.4690 | 6.5573 | 13.5782 |
| 4.0439 | 9.2163 | 3.1380 | 11.5830 | 4.8683 | 19.2469 |
| 2.3534 | 10.1023 | 5.9144 | 2.7443 | 4.3828 | 12.9343 |
| 3.1969 | 2.3832 | 7.4982 | 22.6378 | 3.4239 | 4.1534 |
| 10.7955 | 11.9025 | 9.3565 | 16.5893 | 11.2946 | 24.0416 |
| 40.3177 | 10.2719 | 38.1668 | 25.7490 | 27.6025 | 16.4222 |
| 19.4198 | 20.9311 | 23.5314 | 21.1017 | 18.0436 | 23.6091 |
| 19.7846 | 21.9171 | 19.6500 | 17.2726 | 27.7067 | 28.9641 |
| 26.1448 | 27.2893 | 28.9795 | 23.4561 | 34.9832 | 24.3939 |
| 19.2408 | 35.4537 | 33.8603 | 34.3512 | 25.6864 | 25.2074 |
| 29.7674 | 23.0222 | 19.9895 | 25.0900 | 23.9646 | 22.2552 |
| 26.0507 | 28.7443 | 27.8488 | 31.8092 | 23.3291 | 38.5097 |
| 37.4096 | 28.1230 | 23.5402 | 23.2991 | 39.2026 | 34.8993 |
| 41.9703 | 33.7636 | 29.7331 | 28.9408 | 32.7239 | 47.6175 |
| 24.4182 | 19.2816 | 26.4819 | 25.7852 | 30.8299 | 19.8362 |
| 26.1384 | 31.5121 | 37.7677 | 27.5213 | 17.4678 | 33.2208 |
| 23.9804 | 25.0853 | 35.4589 | 24.5779 | 33.7198 | 27.3530 |
| 21.2638 | 44.9583 | 24.5016 | 23.6719 | 31.4601 | 19.0088 |
| 21.6824 | 19.4404 | 18.9456 | 17.1437 | 35.6352 | 16.0351 |
| 13.5663 | 28.9866 | 17.5990 | 22.6257 | 9.9878 | 8.2797 |
| 5.9969 | 11.4613 | 5.9726 | 4.5866 | 10.3414 | 4.9446 |
| 5.0051 | 4.6294 | 3.5961 | 3.1720 | 2.1722 | 3.9540 |
| 10.5736 | 8.1813 | 15.1029 | 2.7212 | 9.3454 | 13.6903 |
| 8.6685 | 3.1751 | 8.4914 | 4.2121 | 9.3949 | 11.4330 |
| 11.8122 | 11.0833 | 10.2734 | 11.3397 | 8.3568 | 8.8642 |
| **Frames 1-30: Not Talking**<br>**Frames 31-114: Talking**<br>**Frames 115-150: Not Talking** | | | | | |

(a)




(b)

Figure 9. (a) (Left to Right) Variances of Correspondence Distances, (b) Plotted Variances of Correspondence Distances

Experiments were carried out on a number of video sequences. The results were separated by manually separating the sequences to continues, known talking or non-talking clips, and also by separating out video clips that varied through a non-talking, talking and non-talking sequence of status. It was observed that for the former set of video clips the proposed approach achieved an accuracy figure of 97.2% and for the later set 96.1%.

## 4. CONCLUSION

In this paper a novel approach for speaker identification has been described, based on Coherent Point Drift (CPD). Three steps were carried out to achieve the goal, face detection, lip localization and lip movement detection. We have adopted an improved version of a commonly used face detection algorithm in order to obtain the correctly detected faces per frame. A human facial anthropometric division was used in lip localization. We proposed a novel method for lip movement detection based on CPD which was used to compare edge point sets of a given face with the edge point set of a known, non-talking, reference lip. CPD concurrently finds the transformation and the correspondence distance between two different point sets. The CPD method is most effective when approximating still non-rigid transformations and illustrates robust and precise performance regarding noise, outliers and missing points. We carried out experiments to test the proposed system using several video samples; the system attained 97.2% accuracy for the videos which were taken separately for non-talking and talking faces, and 96.1% for the videos which were taken continuously varying through non-talking, talking and non-talking states. We are at present attempting to implement this approach on a practical, widely used, video conferencing system, the *UK access grid.*

## 5. REFERENCES

[1] OpenCVWiki. 2009. FullOpenCVWiki. [Online] (Updated 31 August 2009) Available at: http://opencv.willowgarage.com/wiki/FullOpenCVWiki [Accessed 20 September 2009].

[2] Andriy Myronenko. 2006. Coherent Point Drift (CPD) [Online] (Updated 15 May 2009) Available at: http://www.bme.ogi.edu/~myron/matlab/cpd/ [Accessed 20 September 2009]

[3] Luettin, J., Thacker, N., & Beet, S. (1996). Statistical lip modeling for visual speech recognition. Proceedings of the 8th European Signal Processing Conference (EUSIPCO'96), 10-13.

[4] Tistarelli, M., & Grosso, E. (2000). Active vision-based face authentication. IEEE International Conference on Multimedia and Expo, ICME 2001, 4(18), 299-314.

[5] Sanderson, C., & Paliwal, K. (2004). Identity verification using speech and face information. Digital Signal Processing, 14(5), 449-480.

[6] Hazen, T. J., Weinstein, E., Kabir, R., Park, A., & Heisele, B. (2003). Multi-modal face and speake identification on a handheld device. Paper presented at the Works, Multimodal User Authentication.

[7] Wark, T., & Sridharan, S. (1998). A syntactic approach to automatic lip feature extraction for speaker identification. IEEE International Conference on Acoustics, Speech and Signal Processing. (pp. 3693-3696).

[8] Wark, T., Sridharan, S., & Chandran, V. (1999). Robust speaker verification via fusion of speech and lip modalities. IEEE International Conference on Acoustics, Speech and Signal Processing 1999. ICASSP 99. (pp. 3061-3064).

[9] Dieckmann, U., Plankensteiner, P., & Wagner, T. (1997). Sesam: A biometric person identification system using sensor fusion. Pattern Recognition Letters, 18(9), 827-833.

[10] Frischholz, R. W., & Dieckmann, U. (2000). BioID: A multimodal biometric identification system. IEEE Computer, 33(2), 64-68.

[11] Bigun, J., Duc, B., Fischer, S., Makarov, A., & Smeraldi, F. (1997b). Multi modal person authentication. In H. Wechsler et al., & editor, Nato-Asi advanced study on face recognition, (pp. 26-50).

[12] Nakamura, S. (2001). Fusion of audio-visual information for integrated speech processing. Proceedings Third International Conference on Audio- and Video-Based Biometric Person Authentication: AVBPA 2001 2091, (pp. 127-149).

[13] Shiell, D. J., Terry, L. H., Aleksic, P. S., & Katsaggelos, A. K. (2007, September). An Automated System for Visual Biometrics. Paper presented at the Forty-Fifth Annual Allerton Conference on Communication, Control, and Computing, Urbana-Champaign, IL.

[14] Andriy Myronenko, Xubo Song, and Miguel Carreira-Perpinan. Non-rigid point set registration: Coherent point drift. In B. Sch□olkopf, J. Platt, and T. Ho_man, editors, Advances in Neural Information Processing Systems (NIPS) 19, pages 1009{1016. MITPress, Cambridge, MA, 2007.

[15] A. Myronenko, X. Song, M._A. Carreira-Perpi~n_an: Non-rigid point set registration: Coherent Point Drift, NIPS'19, pages 1009-1016, 2007.