

Real-time Vision Based Mouth Tracking and Parameterization for a Humanoid Imitation Task

Sabri Gurbuz^{a,b}, Naomi Inoue^{a,b} and Gordon Cheng^{c,d}

^a*NICT Cognitive Information Science Laboratories, Kyoto, Japan*

^b*ATR Cognitive Information Science Laboratories, Kyoto, Japan*

^c*ATR-CNS Humanoid Robotics and Computational Neuroscience, Kyoto, Japan*

^d*JST-ICORP Computational Brain Project, Kawaguchi, Saitama, Japan.*

1. Introduction

Robust real-time stereo facial feature tracking is one of the important research topics for a variety multimodal Human-Computer, and human robot Interface applications, including telepresence, face recognition, multimodal voice recognition, and perceptual user interfaces (Moghaddam et al., 1996; Moghaddam et al., 1998; Yehia et al., 1988). Since the motion of a person's facial features and the direction of the gaze is largely related to person's intention and attention, detection of such motions with their 3D real measurement values can be utilized as a natural way of communication for human robot interaction. For example, addition of visual speech information to robot's speech recognizer unit clearly meets at least two practicable criteria: It mimics human visual perception of speech recognition, and it may contain information that is not always present in the acoustic domain (Gurbuz et al., 2001). Another application example is enhancing the social interaction between humans and humanoid agents with robots learning human-like mouth movements from human trainers during speech (Gurbuz et al., 2004; Gurbuz et al., 2005).

The motivation of this research is to develop an algorithm to track the facial features using stereo vision system in real world conditions without using prior training data. We also demonstrate the stereo tracking system through a human to humanoid robot mouth mimicking task. Videre stereo vision hardware and SVS software system are used for implementing the algorithm.

This work is organized as follows. In section 2, related earlier works are described. Section 3 discusses face RIO localization. Section 4 presents the 2D lip contour tracking and its extension to 3D. Experimental results and discussions are presented in Section 5. Conclusion is given in Section 6. Finally, future extension is described in Section 7.

2. Related Work

Most previous approaches to facial feature tracking utilize skin tone based segmentation from single camera exclusively (Yang & Waibel, 1996; Wu et al., 1999; Hsu et al., 2002; Terrillon & Akamatsu, 1999; Chai & Ngan, 1999). However, color information is very sensitive to lighting conditions, and it is very difficult to adapt the skin tone model to a dynamically changing environment in real-time.

Kawato and Tetsutani (2004) proposed a mono camera based eye tracking technique based on six-segmented filter (SSR) which operates on integral images (Viola & Jones, 2001). Support vector machine (SVM) classification is employed to verify pattern between the eyes passed from the SSR filter. This approach is very attractive and fast. However, it doesn't benefit from stereo depth information. Also SVM verification fails when the eyebrows are covered by the hair or when the lighting conditions are significantly different than the SVM training conditions.

Newman et al., (2000) and Matsumoto et al., (1999) proposed to use 3D model fitting technique based on virtual spring for 3D facial feature tracking. In the 3D feature tracking stage each facial feature is assumed to have a small motion between the current frame and the previous one, and the 2D position in the previous frame is utilized to determine the search area in the current frame. The feature images stored in the 3D facial model are used as templates, and the right image is used as a search area firstly. Then this matched image in 2D feature tracking is used as a template in left image. Thus, as a result, 3D coordinates of each facial feature are calculated. This approach requires 3D facial model beforehand. For example, error in selection of a 3D facial model for the user may cause inaccurate tracking results.

Russakoff and Herman (2000) proposed to use stereo vision system for foreground and background segmentation for head tracking. Then, they fit a torso model to the segmented foreground data at each image frame. In this approach, background needs to be modeled first, and then the algorithm selects the largest connected component in the foreground for head tracking.

Although all approaches reported success under broad conditions, the prior knowledge about the user model or requirement of modeling the background creates disadvantage for many practical usages. The proposed work extends these efforts to a universal 3D facial feature tracking system by adopting the six-segmented filter approach Kawato and Tetsutani (2004) for locating the eye candidates in the left image and utilizing the stereo information for verification. The 3D measurements data from the stereo system allows verifying universal properties of the facial features such as convex curvature shape of the nose explicitly while such information is not present in the 2D image data directly. Thus, stereo tracking not only makes tracking possible in 3D, but also makes tracking more robust. We will also describe an online lip color learning algorithm which doesn't require prior knowledge about the user for mouth outer contour tracking in 3D.

3. Face ROI Localization

In general, face tracking approaches are either image based or direct feature search based methods. Image based (top-down) approaches utilize statistical models of skin color pixels to find the face region first, accordingly pre-stored face templates or feature search algorithms are used to match the candidate face regions as in Chiang et al. (2003). Feature based approaches use specialized filters directly such as templates or Gabor filter of different frequencies and orientations to locate the facial features.

Our work falls into the latter category. That is, first we find the eye candidate locations employing the integral image technique and the six segmented rectangular filter (SSR) method with SVM. Then, the similarities of all eye candidates are verified using the stereo system. The convex curvature shape of the nose and first and second derivatives around the nose tip are utilized for the verification. The nose tip is then utilized as a reference for the

selection of the mouth ROI. At the current implementation, the system tracks the person closest to the camera only, but it can be easily extended to a multiple face tracking algorithm.

3.1 Eye Tracking

The pattern of the between the eyes are detected and tracked with updated pattern matching. To cope with scales of faces, various scale down images are considered for the detection, and an appropriate scale is selected according to the distance between the eyes (Kawato and Tetsutani, 2004). The algorithm calculates the intermediate representation of the input image called "Integral image", described in Viola & Jones (2001). Then, a SSR filter is used for fast filtering of bright-dark relations of the eye region in the image. Resulting face candidates around the eyes are further verified by perpendicular relationship of nose curvature shape as well as the physical distance between the eyes, and eye level and nose tip.

3.2 Nose Bridge and Nose Tip Tracking

The human nose has a convex curvature shape and the ridge of the nose from the eye level to the tip of the nose lies on a line as depicted in Fig. 1. Our system utilizes the information in the integral intensity profile of convex curvature shape. The peak of the profile of a segment that satisfies Eqn. 1 using the filter shown in Fig.2 is the convex hull point. A convolution filter with three segments traces the ridge with the center segment greater than the side segments, and the sum of the intensities in all three segments gives a maximum value on the convex hull point. Fig.2 shows an example filter with three segments that traces the convex hull pattern starting from the eye line. The criteria for finding the convex hull point on an integral intensity profile of a row segment is as follows,

$$S_1 < S_2 < S_3 \quad \text{and} \quad \arg\{\max_j(S_1 + 3S_2 + S_3)\}, \quad (1)$$

where S_i denotes the integral value of the intensity of a segment in the maximum filter shown in Fig. 2, and j is the center location of the filter in the current integral intensity profile. The filter is convolved with the integral intensity profile of every row segment. A row segment typically extends over 5 to 10 rows of the face ROI image, and a face ROI image typically contains 20 row segments. Integral intensity profiles of row segments are processed to find their hull points (see Fig.1 using Equation 1 until either the end of the face ROI is reached or until Eqn. 1 is no longer satisfied. For the refinement process, we found that the first derivative of the 3D surface data as well as the first derivative of the intensity at the nose tip are maximum, and the second derivative is zero at the nostril level (Gurbuz et al., 2004a).

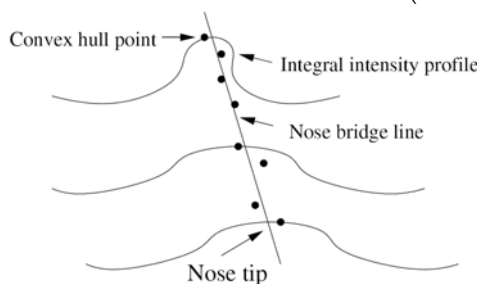


Fig. 1. Nose bridge line using its convex hull points from integral intensity projections.

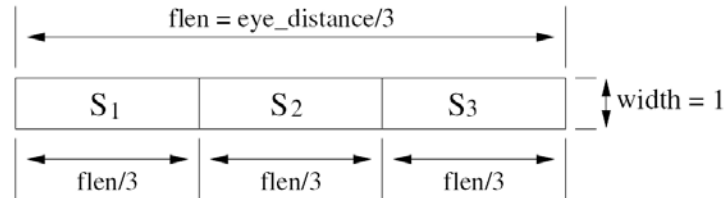


Fig. 2. A three-segment filter for nose bridge tracing.

4. Lip Tracking

The nose tip location is then utilized for the initial mouth ROI selection. Human mouth has dynamic behavior and even dynamic colors as well as presence or absence of tongue and teeth. Therefore, at this stage, maximum-likelihood estimation of class conditional densities for subsets of lip (w_1) and non-lip (w_2) classes are formed in real-time for the Bayes decision rule from the left camera image. That is, multivariate class conditional Gaussian density parameters are estimated for every image frame using an unsupervised maximum-likelihood estimation method.

4.1 Online Learning and Extraction of Lip and Non-lip Data Samples

In order to alleviate the influence of ambient lighting on the sample class data, chromatic color transformation is adopted for color representation (Chiang et al., 2003; Yang et al., 1998). It was pointed out (Yang et al., 1998) that human skin colors are less variant in the chromatic color space than the RGB color space. Although in general the skin-color distribution of each individual may be modeled by a multivariate normal distribution, the parameters of the distribution for different people and different lighting conditions are significantly different. Therefore, online learning and sample data extraction are important keys for handling different skin-tone colors and lighting changes. To solve these two issues, the authors proposed an adaptation approach to transform the previous developed color model into the new environment by combination of known parameters from the previous frames. This approach has two drawbacks in general. First, it requires an initial model to start, and second, it may fail in the case of a different user with completely different skin-tone color starts using the system.

We propose an online learning approach to extract sample data for lip and non-lip classes to estimate their distribution in real time. Chiang et al. (2003) in their work provides hints for this approach. They pointed out that lip colors are distributed at the lower range of green channel in the (r,g) plane. Fig. 4 shows an example distribution of lip and non-lip colors in the normalized (r,g) space.

Utilizing the nose tip, time dependent (r,g) spaces for lip and non-lip are estimated for every frame by allowing \mathcal{E} % (typically 10%) of the non-lip points stay within the lip (r,g) space as shown in Fig. 4. Then, using the obtained (r,g) space information in the initial classification, the pixels below the nostril line that falls within the lip space are considered as lip pixels, and the other pixels are considered as non-lip pixels in the sample data set extraction process, and RGB color values of pixels are stored as class attributes, respectively.



Fig. 3. Left image: result of the Bayes decision rule, its vertical projection (bottom) and integral projection of intensity plane between nose and chin (right). Middle image: estimated outer lip contour using the result of the Bayes rule. Right image: a parameterized outer lip contour.

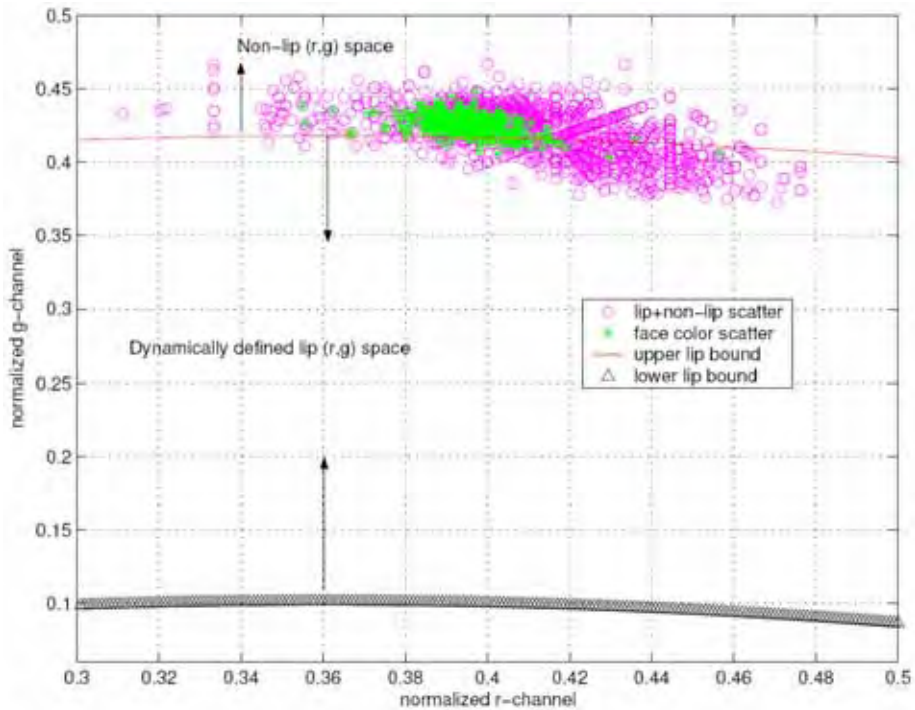


Fig. 4. Dynamically defined lip and non-lip (r,g) spaces.

In most cases, sample data contains high variance and it is preferable to separate the data into subsets according to its time dependent intensity average. Let avg_L and D_k be the intensity average and k^{th} subset of the lip class, respectively. The subsets of the lip class are separated according to lip class' intensity average as

$$\begin{cases} \text{assign to } D_1 & \text{if } x_{intensity} < avg_L, \\ \text{assign to } D_2 & \text{if } avg_L/2 < x_{intensity} < 3avg_L/2, \\ \text{assign to } D_3 & \text{if } x_{intensity} > avg_L. \end{cases} \quad (2)$$

Using the same concept in Eqn. 2, we also separate the non-lip data samples into subsets according to intensity average of the non-lip class. Fig. 5 depicts simplified conditional density plots in 1D for the subsets of an assumed non-lip class.

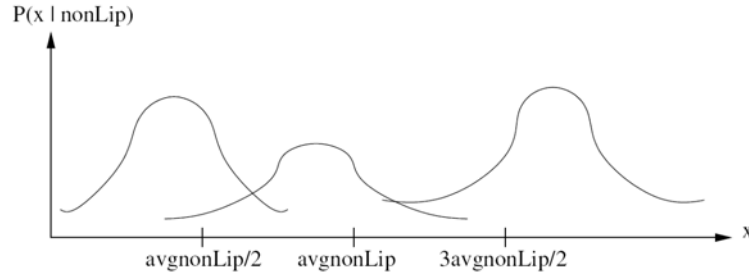


Fig. 5. Example class conditional densities for subsets of non-lip class.

4.2 Maximum-Likelihood Estimation of Class Conditional Multivariate Normal Densities

The mean vector and covariance matrix are the sufficient statistics to completely describe a distribution of the normal density. We utilize a maximum-likelihood estimation method for the estimation of a class conditional multivariate normal density described by

$$p(\mathbf{x}|i) = \frac{1}{\sqrt{(2\pi)^n \|\Sigma_i\|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)\Sigma_i^{-1}(\mathbf{x} - \mu_i)^T\right\}, \quad (3)$$

where i may be w_1 , or w_2 , or subset of a class. $\mu_i = E[x]$ is the mean value of the i^{th} class. Σ_i is the $n \times n$ (in this work, n is number of color attributes so $n = 3$) covariance matrix defined as

$$\Sigma_i = E[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T] \quad (4)$$

where $\|\cdot\|$ represents the determinant operation, and $E[\cdot]$ represents the expected value of a random variable. Unbiased estimates of the parameters μ_i and Σ_i are estimated by using the sample mean and sample covariance matrix.

4.3 Bayes Decision Rule

Let \mathbf{x} be an observation vector formed from RGB attributes of a pixel location in an image frame. Our goal is to design a Bayes classifier to determine whether \mathbf{x} belongs to w_1 or w_2 in two class classification problem. The Bayes test using a posteriori probabilities may be written as follows:

$$p(w_1|\mathbf{x}) \stackrel{w_1}{\geq} p(w_2|\mathbf{x}), \quad (5)$$

where $p(w_i | x)$ is the a posteriori probability of w_i given x . Equation (5) shows that if the probability of w_1 given x is larger than the probability of w_2 , then x is declared belonging to w_1 , and vice versa. Since direct calculation of $p(w_i | x)$ is not practical, we can re-write the

a posteriori probability of w_i using Bayes' Theorem in terms of a priori probability and the conditional density function $p(x|w_i)$, as

$$p(w_i|\mathbf{x}) = \frac{p(\mathbf{x}|w_i)p(w_i)}{p(\mathbf{x})} \quad (6)$$

where $p(x)$ is the density function and is positive constant for all classes. Then, re-arranging both sides, we obtain

$$L(\mathbf{x}) = \frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} \stackrel{w_1}{\geq} \frac{p(w_2)}{p(w_1)} \quad (7)$$

where $L(x)$ is called the likelihood ratio, and $p(w_1)/p(w_2)$ is called the threshold value of the likelihood ratio for the decision. Because of the exponential form of the densities involved in Equation (7), it is preferable to work with the monotonic discriminant functions obtained by taking the logarithm as follows.

$$q_{w_i}(\mathbf{x}) = \ln(p(\mathbf{x}|w_i)p(w_i)), \quad (8)$$

thus, by re-arranging Equation (8), we get

$$q_{w_i}(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)\Sigma_i^{-1}(\mathbf{x} - \mu_i)^T + c_i \quad (9)$$

Where

$$c_i = \ln p(w_i) - (1/2) \ln 2\pi - (1/2) \|\Sigma_i\|$$

is a constant for this image frame. In general, Equation (9) has only nonlinear quadratic form and a summation, and using this equation, the Bayes rule can be implemented for real-time lip tracking as follows.

$$q_{w_1}^*(\mathbf{x}) \stackrel{w_1}{\geq} \underset{w_2}{q_{w_2}^*(\mathbf{x})}, \quad (10)$$

Where

$$q_i^*(\mathbf{x}) = \max\{q_i^{(1)}(\mathbf{x}), q_i^{(2)}(\mathbf{x}), q_i^{(3)}(\mathbf{x})\}$$

for $i = \{w_1, w_2\}$ and referring to Fig. 5. Threshold value of the likelihood ratio as shown in Eqn. (7) is based on a priori class probabilities. In our implementation, equally likely a priori class probabilities are assumed.

4.4 Mouth Outer Contour Parameterization in 2D

After mouth tracking algorithm locates the mouth region, outer lip contours of the speaker's lips in left camera image are detected (see Fig. 3). Then, the outer contour as a whole is parameterized by a generalized ellipse shape which is obtained using the estimated outer contour data. A parametric contour is found that corresponds to the general quadratic equation $a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + a_6 = 0$, where a_i s are constants, and a_1 and a_3 are non-zero. Let us denote the 2D positions over the traced outer lip contour as

$$\begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_N \\ y_1 & y_2 & y_3 & \dots & y_N \end{bmatrix}. \quad (10)$$

The basic form used in the elliptical parameter estimation in matrix notation is $Ma = 0$ where $a = [a_1 \dots a_6]^T$. The dimensionality of M is the number of points, N , in the segment multiplied by 6 (that is, $N \times 6$). Each row of M corresponds to one point in the segment. The

parameters of each contour are then solved using the least-squares method to find a_i s, where $i=1,2,\dots,6$.

Using the estimated parameters, parametric lip contour data can be re-generated for each image frame. Five points are sufficient to represent a general elliptical shape, leading to a significant data reduction and representation.



Fig. 6. Screen capture of tracked outer lip contours for various skin tone colors and different lighting conditions.

4.5 Estimation of 3D Mouth Outer Contour

Once the outer lip contour points of the speaker's lips in left camera image are found then their stereo disparity values from the right image can be calculated utilizing the previously found horopter information. Fig. 7 shows stereo and disparity images. Knowing a pixel location (x,y) in the left camera image and its disparity in the right camera image, we can calculate its 3D (X,Y,Z) coordinates with respect to the camera coordinate system as shown in Fig. 8.



Fig. 7. Screen capture of the left and right camera images, and their disparity map around the face region of interest.



Fig. 8. Screen capture of the left and right camera images, and OpenGL plot of texture mapped 3D face reconstructed by the stereo algorithm.

5. Experimental Results and Discussion

In this paper, our work focused on a real-time stereo facial feature tracking algorithm. Intensity information is used for initial eye candidates in the left image. Relationship of the eyes and the nose are verified using 3D data. Then, the nose tip is utilized as a reference point for mouth ROI. RGB color attributes of lip pixels are used for the lip tracking.

The proposed stereo facial feature tracking algorithm has been tested on live data from various users without using any special markers or paintings. Fig. 6 shows tracked lip contour results for various users under various lighting conditions. The stereo facial feature tracking algorithm which utilizes Videre stereo hardware works around 20 frames per second (un-optimized) on a 2 GHz notebook PC under Windows platform. We also demonstrate the developed stereo tracking system through a human to humanoid robot mouth mimicking task. The humanoid head (Infanoid) is controlled via serial communication connected to a PC. Commands to the mouth is send every 50 ms (20 hz), at the same rate as the processing of facial features tracking - thus, allowing real-time interaction. The communication between the vision system and the control PC is via a 1Gbit Ethernet network. The opening and closing of the person's mouth is directly mapped to the mouth of the humanoid's mouth, with a simple geometric transform. The Infanoid robot head is shown in Fig. 10 which was developed by Kozima (2000).

6. Conclusions

A new method for stereo facial feature tracking of individuals in real world conditions is described. Specifically, stereo face verification and an unsupervised online parameter estimation algorithm for the Bayesian rule is proposed. That is, a speaker's lip colors are learned from the current image frame using the nose tip as a reference point. Vertical and horizontal integral projections are utilized to guide the algorithm in picking out the correct lip contour. In the final stage, estimated outer contour data for every image frame of the left camera is parameterized as a generalized ellipse. Then, utilizing the contour pixel locations in the left camera image and their disparity of the right camera image, we calculate their 3D (X,Y,Z) coordinates with respect to the camera coordinate system. Future work for vision includes extraction of 3D coordinates of other facial points such as eyebrows, chin and cheek, and further extending the work to the multiple face tracking algorithm.

7. Humanoid Robotics Future Extension

Future extensions of this work include developing a machine learning method for smooth mouth movement behavior to enable humanoids to learn visual articulatory motor tapes for any language with minimal human intervention. Fig. 9 shows the flow diagram of the system. Such a system should extract and store motor tapes by analyzing a human speaker's audio-visual speech data recorded from predetermined phonetically-balanced spoken text to create a mapping between the sound units and the time series of the mouth movement parameters representing the mouth movement trajectories. These motor tapes can then be executed with the same time index of the audio, yielding biologically valid mouth movements during audio synthesis.

We call the system the *text-to-visual speech (TTVS) synthesis* system. It can be combined with a concatenative speech synthesis system, such as Festival (Black & Taylor, 1997; Sethy & Narayanan, 2002; Chang et al., 2000) to create a text-to-audiovisual speech synthesis system for humanoids.

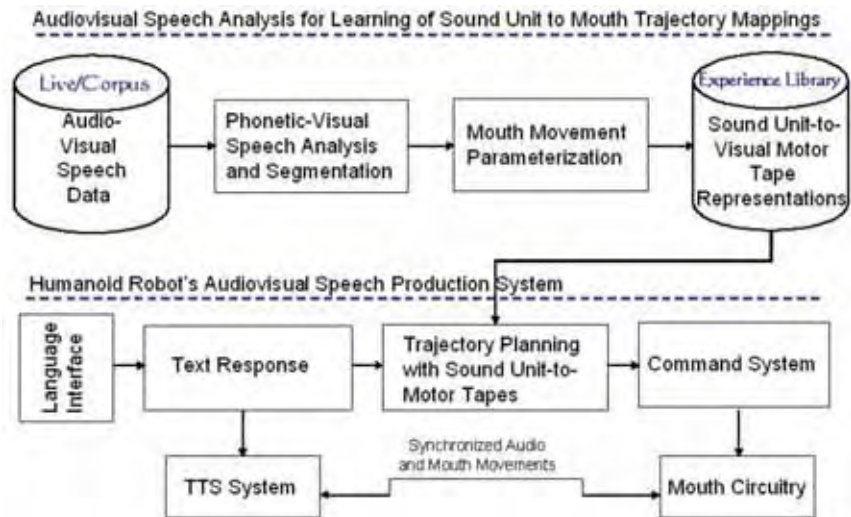


Fig. 9. Future extension to a TTS based speech articulation system for Humanoids.

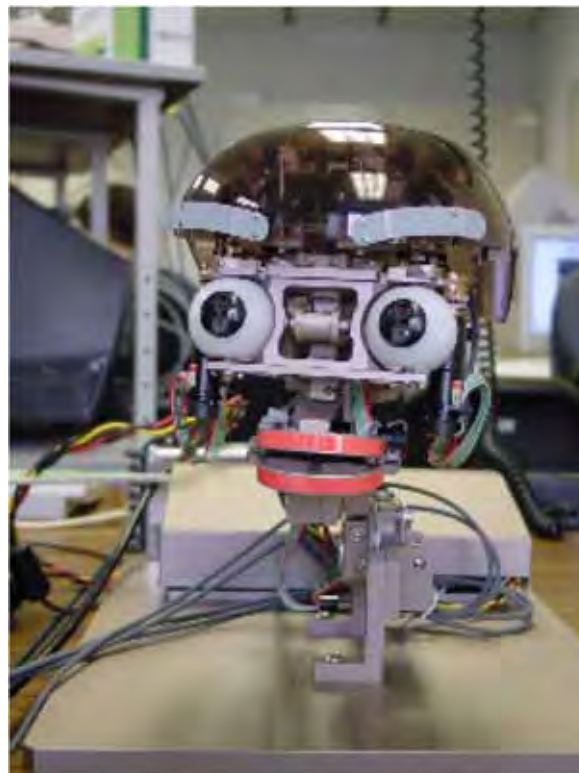


Fig. 10. Infantoid robot utilized for human to humanoid robot mouth imitation task.

A concatenative synthesis system creates indexed waveforms by concatenating parts (diphones) of natural speech recorded from humans. Using the same concatenative synthesis concept, the proposed TTVS system can concatenate corresponding mouth movement primitives. Thus, the system is capable of generating sequences of entirely novel visual speech parameters that represent the mouth movement trajectories of the spoken text. A humanoid agent equipped with TTS and TTVS systems can produce novel utterances, and so is not limited to those recorded in the original audio-visual speech corpus. With these capabilities, the humanoid robot can robustly emulate a person's audiovisual speech. A detailed explanation of this extension is described in Gurbuz et al. (2004b). Also, we will extend the work to include imitation of other facial movements, as the vision system expands to stereo and track additional features such as eyebrows, and perform perception studies to ascertain the effect of more accurate speech and face movement cues on naturalness and perceptibility in humanoids.

8. Acknowledgment

This research was conducted as part of "Research on Human Communication" with funding from the National Institute of Information and Communications Technology (NiCT), Japan. Thanks are due to Dr. Hideki Kozima for the use of his Infanoid robot and Shinjiro Kawato for the original eye tracking work extended in this paper.

9. References

- Black, A., & Taylor, P. (1997). *The Festival Speech Synthesis System*. University of Edinburgh.
- Chai, D. & Ngan, K. N., (1999). Face segmentation using skin-color map in videophone applications. *IEEE Trans. on Circuits and Systems for Video Technology* 9 (4), 551-564.
- Chang, S., Shari, L. & Greenberg, S. (2000). Automatic phonetic transcription of spontaneous speech (American English), *International Conference on Spoken Language Processing, Beijing, China*.
- Chiang, C. C., Tai, W. K., Yang, M. T., Huang, Y. T. & Huang, C. J., (2003). A novel method for detecting lips, eyes and faces in real-time. *Real-Time Imaging* 9, 277-287.
- Gurbuz, S., Shimizu, T. & Cheng, G. (2005). Real-time stereo facial feature tracking: Mimicking human mouth movement on a humanoid robot head. *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2005)*.
- Gurbuz, S., Kinoshita, K., & Kawato, S., (2004a). Real-time human nose bridge tracking in presence of geometry and illumination changes. *Second International Workshop on Man-Machine Symbiotic Systems, Kyoto, Japan*.
- Gurbuz, S., Kinoshita, K., Riley, M. & Yano, S., (2004b). Biologically valid jaw movements for talking humanoid robots. *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2004), Los Angeles, CA, USA*.
- Gurbuz, S., Tufekci, Z., Patterson, E. & Gowdy, J., (2001). Application of affine invariant Fourier descriptors to lipreading for audio-visual speech recognition. *Proceedings of ICASSP*.
- Hsu, R. L., Abdel-Mottaleb, M. & Jain, A. K., (2002). Face detection in color images. *IEEE Trans. on PAMI* 24 (5), 696-706.
- Kawato, S. & Tetsutani, N., (2004). Scale adaptive face detection and tracking in real time with SSR filter and support vector machine. *Proc. of ACCV*, vol. 1.

- Kozima, H., (2000). NICT infanoid: An experimental tool for developmental psychorobotics. International Workshop on Developmental Study, Tokyo.
- Matsumoto, Y. & Zelinsky, A., (1999). Real-time face tracking system for human robot interaction. Proceedings of 1999 IEEE International Conference on Systems, Man, and Cybernetics (SMC'99). pp. 830-835.
- Moghaddam, B., Nastar, C. & Pentland, A., (1996). Bayesian face recognition using deformable intensity surfaces. In: IEEE Conf. on Computer Vision and Pattern Recognition.
- Moghaddam, B., Wahid, W. & Pentland, A., (1998). Beyond eigenfaces: Probabilistic matching for face recognition. In: International Conference on Automatic Face and Gesture Recognition.
- Newman, R., Matsumoto, Y., Rougeaux, S. & Zelinsky, A., (2000). Real-time stereo tracking for head pose and gaze estimation. Proceedings. Fourth IEEE International Conference on Automatic Face and Gesture Recognition.
- NICT-Japan, Infanoid project. <http://www2.nict.go.jp/jt/a134/infanoid/robot-eng.html>.
- Russako, D. & Herman, M., (2000). Head tracking using stereo. Fifth IEEE Workshop on Applications of Computer Vision.
- Sethy, A. & Narayanan, S. (2002). Refined speech segmentation for concatenative speech synthesis, *International Conference on Spoken Language Processing, Denver, Colorado*.
- Terrillon, J. C. & Akamatsu, S., (1999). Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. Proc. 12th Conf. on Vision Interface. pp. 180-187.
- Viola, P. & Jones, M., (2001). Robust real-time object detection. Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, Vancouver, Canada.
- Wu, H., Cheng, Q. & Yachida, M., (1999). Face detection from color images using a fuzzy pattern matching method. IEEE Trans. on PAMI 21 (6), 557-563.
- Yang, J., Stiefelhagen, R., Meier, U. & Waibel, A., (1998). Visual tracking for multimodal human computer interaction. Proceedings of the SIGCHI conference on Human factors in computing systems.
- Yang, J. & Waibel, A., (1996). A real-time face tracker. In: Proc. 3rd IEEE Workshop on Application of Computer Vision. pp. 142-147.
- Yehia, H., Rubin, P.E., & Vatikiotis-Bateson, E. (1998). "Quantitative association of vocal tract and facial behavior," *Speech Communication*, no. 26, pp. 23-44.