USING LOCAL BINARY PATTERNS FOR NON-CONTACT OPTICAL TONGUE TRACKING

Ahmad Ghadiri, James R. Green, and Andrew E. Marble
Department of Systems and Computer Engineering, Carleton University

ABSTRACT

In this study, a real-time tongue tracking system is developed. The goal is to track a user's tongue in a safe, non-contact manner using a webcam and image processing algorithms. This system functions in a two-level architecture. First, it detects the approximate location of the mouth. Then, exact mouth state and tongue direction are determined. A rapid object detection algorithm which makes use of Multi-scale Block Local Binary Patterns (MB-LBP) is applied in this system. Instead of pixel intensities, this algorithm employs MB-LBP features for computations in processing digital images which significantly reduces computational requirements and increases the frame rate. The Gentle Adaptive Boosting meta-algorithm (AdaBoost) is used to obtain a classifier for each mouth/tongue state. Six-state classification accuracy is measured using a hold-out test with six subjects of varying ethnicities. Accuracy is comparable with that of our previous prototype, but is more robust to ambient lighting and head pose. Accuracy is expected to further increase with the collection of more training data.

INTRODUCTION

A wide variety of patient groups benefit from speech therapy, including those with disabilities or those recovering from stroke. Non-speech motor oral exercises have shown promise when included within a speech therapy regime [1]. Physical exercise of the tongue is frequently used to increase its dexterity, ultimately leading to greater control [2]. This system will guide users in performing non-speech motor oral exercise with their mouth and tongue to make the therapy more entertaining and engaging than the current speech therapy exercises.

While existing systems require physical instrumentation of the tongue [3], or the use of a dedicated ultrasound device [4], to detect the current state of the mouth and tongue and return feedback from movements of tongue, our system makes use of a non-invasive approach combining a webcam and advanced image processing algorithms. After detecting the location of the subject's face and mouth in each video frame, it is determined whether the mouth is closed or open. When the mouth is open, further analysis is done to categorize the direction of the tongue (up, down, right, left). The output of this component can be used to develop an amusing interface for the user. A prototype of our system was reported previously [5]. This system employed hard thresholds on the red, green, and blue color channels and therefore system performance degraded quickly as ambient light levels varied from the ideal training conditions. This study reports on significant advances in the image processing algorithms employed within our system. This system has shown to be more accurate and also faster when compared to our previous tongue tracking system. Furthermore, it can handle slight head rotations and variations in ambient lighting. This robustness is important, given that one of the targeted user groups are children who may be less apt to remain still during the exercises.

METHODS

This system is composed of a two-level hierarchy of classifiers. First, a robust face detection component is implemented to track and segment the user's face. A region of interest is derived from the face location (i.e. lower half of the face region in image). Additionally, six classifiers, each corresponding to a mouth/tongue state, are employed in parallel to distinguish different expressions of mouth and tongue. An ideal open mouth classification using this method is shown in figure 1, where the output from the "Open Mouth Detector" would have greater magnitude than all other detectors.

Face Detection

In order to determine of the location of mouth, the system firstly detects the location of the face in the current frame. Haar-like feature based classifiers have shown significant accuracy and speed for locating the face in real-time applications [6]. Furthermore, an extended set of Haar-like features has been introduced in the literature which appears to have fewer false alarms and a higher detection rate as it covers more positions despite employing the same number of training images [7]. Such features result in a

simpler image representation compared with raw pixel intensities. This leads to an increase in accuracy and a reduction in computational complexity for real-time applications. These features are fed as input to a machine learning algorithm to produce a classifier. More specifically, in this method, the Adaptive boosting meta-algorithm (AdaBoost) [6] is applied to generate a filter chain of weak classifiers which, taken together, forms a sturdy cascade of classifiers.

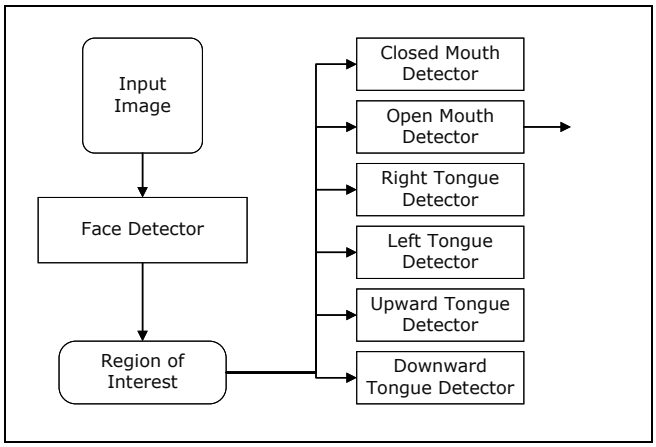

Figure 1: System diagram for "Open Mouth" case

A representation of how several classifiers come together to form a cascade and the overall functioning of the sturdy cascade is shown in figure 2. This algorithm searches for rectangular Haar-features, a huge number of which are found in an image sub-window. Initial classifiers (i.e. h_1, h_2…) are trained to quickly recognize and rule-out a large number of negative sub-windows that does not contain features of the object. These classifiers are not expected to have a very low error rate. Thus with very little processing, they will act rapidly and reject a massive set of irrelevant rectangular features found in the image. In other words instead of passing all the features to a single but strong classifier, as the input proceeds through weak classifiers, the number of features will be reduced and then the remaining "difficult to classify" data will be fed to classifiers with higher precision. In this system, a trained cascade is implemented for real-time face detection. This cascade has been shown to be resilient to scale, light variations and rotations of the head [6].

Mouth State and Tongue Direction Detection

Once a face is detected within a frame, all regions falling outside of the face or in the upper half of the face can be discarded as they do not correspond to our region of interest in this study.

Therefore, at this state, an approximate location of the mouth is determined. On account of the fact that there are no existing classifiers trained for mouth and tongue

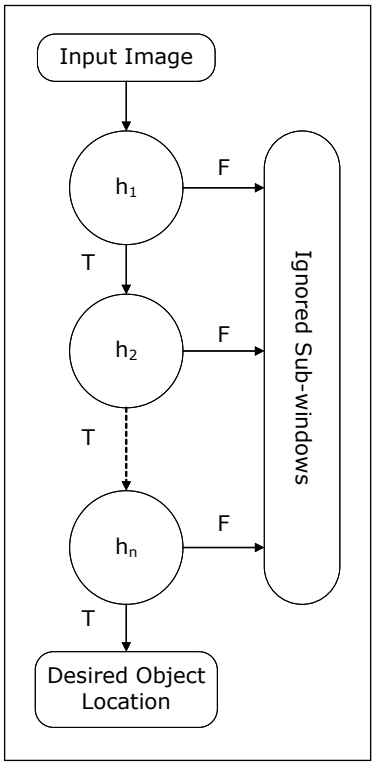

Figure 2: Illustration of cascaded classifiers

state determination, several boosted cascades of classifiers have been trained to distinguish the states of mouth and tongue direction. These cascades are conceptually similar to the cascade used for face detection, except that they make use of a novel image representation scheme.

There has been recent interest in image representation and texture description using Local Binary Patterns (LBP) [8]. The LBP operator allocates a binary number to every pixel by thresholding a 3x3-neighborhood. Starting at the top-left of the 3x3 neighborhood, each pixel is compared with the center pixel, and is assigned a value of 1 if it has greater (grey scale) intensity than the center pixel and 0 otherwise. Therefore a 3x3-neighborhood will generate an 8 bit binary number. An example computation for this representation is shown in figure 3. A computed histogram of the labels (i.e. binary numbers) can be used as a texture descriptor which is robust to scale since it can simply be searched at different scales within an image [8, 9]. However, LBP features may sometimes fail to capture large scale structures as they are associated only with a 3x3 neighborhood. To address this issue, it has been suggested that using a

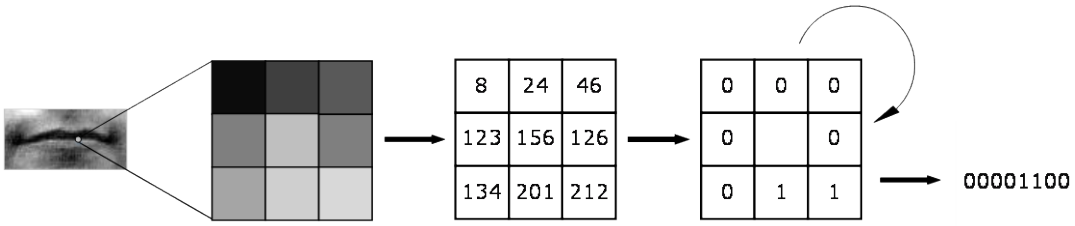

Figure 3: Sample calculation of local binary pattern for a 3x3 pixel neighborhood

larger pixel neighborhood [9], for example (8,1), (16,2) or (8,2) where each tuple takes the form of (number of pixels, radius of circle), or considering blocks of pixels instead of individual pixels [10] will effectively capture larger structures. Multi-scale Block Local Binary Patterns are considered a powerful technique for detecting complex objects in real-time applications [10]. In fact, labeling is done based on the average value of blocks of pixels. For instance, by considering a 4x4 pixel block, computing the average of each block, and then comparing it to adjacent blocks in a 3x3 neighborhood of the same size blocks, a binary number containing more information will be produced which can be used to capture larger structures appearing in an image. After extracting MB-LBP features, histograms of labels are computed. Finally, they are similarly passed as input through Gentle Adaptive Boosting machine learning algorithm to train classifiers each of which is capable of distinguishing one mouth or tongue state from all of other possible states. To increase the detection rate in dark environments, prior to computing the MB-LBP features, histogram equalization [11] is applied to the frame so that the system will function properly in such environments. Subsequent to locating the face, mouth and tongue detection is performed by a number of cascades of boosted classifiers based on MB-LBP features.

RESULTS

In this study, because of the unavailability of mouth and tongue position data, a new data set has been gathered. This data set contains a total of 11,177 facial images of 6 subjects taken from 6 mouth/tongue states. The 6 subjects represented diverse ethnicities and skin tones. Images are categorized into 6 classes of closed/open mouth and right/left/down/up tongue directions. During data collection, subjects were asked to slowly rotate their head left, right, up and down within 60 degrees from the plane of the camera. Subjects were also asked to lean right and left in order to produce images exhibiting head angles from 0 to 45 degrees. In our experiments, we have used the Open Source Computer Vision (OpenCV) library [12, 13]. Ethical clearance was obtained from the Carleton University Research Ethics Board prior to data collection.

Each class (i.e. mouth or tongue state) of this data set is used to train a distinct MB-LBP cascade. Each classifier is trained to recognize one of the 6 mouth/tongue states. Since there are no existing classifiers to detect such objects in an image, mouth regions were manually cropped and scaled prior to training the binary classifiers. Given that all classifiers are functioning at the same time, it is crucial to keep a low false alarm rate. Mouth regions are rigorously cropped, and an approximate initial image size of 22x20 or 40x20 for training images is considered so that they will result in training cascades with minimum false-positive rates [14]. Moreover, classifiers were trained in 18 to 20 stages. In this study, images of 5 subjects were used for training, and images of the sixth subject were used in a hold-out test to determine the accuracy of the trained cascade classifiers. Approximately 700 negative (i.e. background samples) and 1534, 1077, 1971, 1616, 1453 and 1233 positive samples of closed, open, right, left, down, up positions respectively were used for training each cascade.

Table 1: 6-state classification Performance

Cascade	Actual Class/Number of Testing Frames					
	Close	Open	Right	Left	Down	Up
	317	311	370	221	432	402
Closed	**245**	1	5	2	0	1
Open	0	**275**	0	0	0	0
Right	0	119	**369**	6	12	0
Left	0	62	0	**221**	13	0
Down	0	2	88	3	**350**	2
Up	1	11	122	10	0	**401**

Table 1 represents the detailed classification results of each cascade. As mentioned before, a total number of 9124 images corresponding to 5 subjects with 6 mouth/tongue states were used to train six MB-LBP boosted cascades of classifiers. Each of the cascades is capable of detecting one specific state. In the testing phase, 2053 images related to a held-out test subject having the same 6 mouth/tongue states were used to determine the accuracy and false-positive rate of this system. In the results table, diagonal cells contain the number of true-positive classifications, while other

cells in the table represent classification errors. Each row includes the functionality details of one cascade on frames of all states. As an illustration for closed cascade, 245 out of 317 frames with closed mouth were correctly detected as closed and the rest were missed. Using the same cascade, only 1 out of 311 frames with open mouth was incorrectly classified as closed mouth while other frames containing an open mouth (correctly) resulted in a negative cascade output. In summary, the 6-way classifier had an accuracy of 91.03%. Considering that 60 % of the test images included head turn and tilt (i.e. that the subject was not constrained to perfectly face the camera), this level of accuracy far exceeds our initial prototype system [5]. Table 2 demonstrates the accuracy as well as false-positive rate of each binary MB-LBP boosted cascade tested on a single subject. An overall accuracy of 91.03 % and false-positive rate of 4.57 % has been achieved.

Table 2: Overall accuracy and false positive rate

Cascade	Accuracy (%)	False Positive Rate (%)
Closed	77.28	0.52
Open	88.42	0.00
Right	99.72	8.18
Left	100.00	4.09
Down	81.01	5.89
Up	99.75	8.77
Overall	**91.03**	**4.57**

To train and test our classifiers, we have used a desktop computer with a 2.4GHz Intel Core 2 Duo CPU and 2GB of RAM. Face detection took approximately 35-55 ms per frame using a 640x320 pixel webcam. After face detection, each MB-LBP mouth/tongue state classifier required 4-6 ms. Hence, an average time of 70 ms is required per frame for detecting both face and mouth/tongue state. This time is compatible with real-time operation, allowing for almost 15 frames per second.

CONCLUSION AND FUTURE WORK

The fast processing time associated with boosted cascades of classifiers makes them ideal in real-time applications [6]. Recently, MB-LBP features have been introduced in the literature which are used to implement a new simple yet effective texture descriptor [10]. In this study, we have developed 6 classifiers using the aforementioned methodology on a relatively small training data set. The classifiers have shown an approximate accuracy of 91%. The accuracy of this approach is expected to further improve as more training data can be gathered representing a greater diversity of mouth/tongue shapes, ambient lighting, and skin tone. Additionally, the number of mouth/tongue states can be increased if required, given more training data. Although our experiments do not contain all possible inputs such as images from people of all possible ethnicities and skin tones, data can be collected to further improve this aspect of the system. This approach can serve as the basis for a robust, non-invasive, and real-time speech therapy system.

REFERENCES

[1] R.J. McCauley, E. Strand, G.L. Lof, "Evidence-Based Systematic Review: Effects of Nonspeech Oral Motor Exercises on Speech", American Journal of Speech-Language Pathology, 2009, pp. 343-360

[2] H.M. Clark, "The Role of Strength Training In Speech Sound Disorders", Seminars In Speech And Language, 2008, pp. 276-283

[3] M. Aron, M. Berger, E Kerrien, Y. Laprie, "Coupling Electromagnetic Sensors and Ultrasound Images for Tongue Tracking", 7th International Seminar on Speech Production, 2006

[4] M. Stone, "A Guide To Analysing Tongue Motion From Ultrasound Images", Clinical linguistics & phonetics, 2005, pp. 455-501

[5] K. Mulligan, J. LaRocque, J.R. Green, "A Low Cost Non-Contact Approach to Tongue Tracking for Special Needs Children", Canadian Medical and Biological Engineering Conference (CMBEC), 2007

[6] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 511-518

[7] R. Lienhart, J. Maydt. "An Extended Set of Haar-like Features for Rapid Object Detection", IEEE International Conference on Image Processing (ICIP), 2002, pp. 900-903

[8] T. Ojala, M. Pietikainen, T. Maenpaa, "Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns", European Conference on Computer Vision (ECCV), 2000, pp 404-420

[9] T. Ahonen, A. Hadid, M.I. Pietikainen, "Face Recognition with Local Binary Patterns", European Conference on Computer Vision (ECCV), 2004, pp. 469-481

[10] S. Liao, X. Zhu, Z. Lei, L. Zhang and S. Z. Li. "Learning Multi-scale Block Local Binary Patterns for Face Recognition", International Conference on Biometrics (ICB), 2007, pp. 828-837.

[11] R. Hummel, "Image Enhancement by Histogram Transformation", Computer Graphics and Image Processing, 1977, pp. 184–195

[12] G. Bradski, A. Kaehler, "Learning OpenCV: Computer Vision with the OpenCV Library", 2008

[13] R. Lagani Re, "OpenCV 2 Computer Vision Application Programming", 2011

[14] R. Lienhart, A. Kuranov, V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection", Pattern Recognition Lecture Notes in Computer Science, 2003, pp. 297-304