Adaptive farbbasierte Mundsegmentierung

A. Panning, A. Al-Hamadi und B. Michaelis

Otto-von-Guericke Universität Magdeburg
{Axel.Panning, Ayoub.Al-Hamadi}@ovgu.de

Die Segmentierung von Mund und Lippen sind ein fundamentales Problem der Gesichtsausdrucksanalyse. In diesem Beitrag zeigen wir eine Methode zur Segmentierung, basierend auf einem Histogramm der Mund-Region of Interest die in den $rg \rightarrow g$ Farbkanal transformiert wurde. Diese Farbtransformation stellte sich gegenüber anderen Farbtransformationen als optimal heraus. Ausgehend von einer initialen Schätzung wird die Verteilung des Hautfarbmodells approximiert, mit dessen Hilfe die initiale Schätzung verfeinert und somit der ideale Schwellwert für die Binarisierung bestimmt wird. In den Experimenten zeigte sich, dass die beschriebene Methode performanter und adaptiver ist als bisherige rein farbbasierte Verfahren.

1 Einführung

Im Rahmen der automatischen Mimikerkennung spielt die Segmentierung des Mundes eine wichtige Rolle. Während Augen und Augenbrauen für bewusste Mimik eingesetzt werden, vollzieht der Mund häufig auch unbewusste Mimiken. Das finale Problem ist die Binarisierung in Haut- und Lippenpixel. Genau dieser Schritt soll im vorliegenden Beitrag betrachtet werden. Je nach Anwendungen werden unterscheiden sich die Anforderungen bezüglich Genauigkeit und Robustheit. Viele bekannte Ansätze sind spezialisiert auf die Auswertung von Farbbildern und zielen dabei auf die Farbdivergenz zwischen Mund und Haut ab [2–5, 9, 10]. Dabei wird initial der RGB Farbraum in eine einkanalige Intensitätsmap transformiert, in der sich die Lippenpixel des Mundes möglichst gut vom Rest der Haut abheben. Das meiste Gewicht erhalten dabei oft der Rot- sowie der Grünkanal des RGB. Daneben gibt es auch viele Arbeiten, die Schwarzweissbilder verwenden und dem entsprechend nicht auf Farbinformationen zurückgreifen können [6, 8, 12]. Grundsätzlich gibt es zwei Methoden für die Extraktion des Mundes. Die erste zielt vornehmlich auf die Detektion der Gradienten ab, die beim Übergang von Lippen zur Haut erwartet werden [2, 4]. Sie benutzen dabei Aktive Konturmodelle (*Active Shape Models - ASM*) [4] oder deformierbare Templates [2]. Einige Ansätze stabilisieren die Konturmodelle durch unterstützende Trackingpunkte, um ein abdriften des Modells zu verhindern [1, 6]. Die grundlegende Annahme, dass an den

Übergängen von Lippen zur Haut starke Gradienten entstehen kann jedoch zu Problemen führen. In monochromen Bildern können bereits minimale Schatteneffekte (tritt hauptsächlich an den Unterlippen auf) zu ernsthaften Störungen führen. Farbbilder und ihre transformierte Intensitätsmap können hier Abhilfe verschaffen. Dennoch ist der Kantenübergang oft nicht signifikant genug. Im Bereich des Mundes auf den Lippen selbst entstehen oft stärkere Gradienten (Glanzpunkte, Lippeneigentextur, innere Mundkanten bei geöffnetem Mund usw.) als an der Mundaußenkante. Die zweite Grundlegende Methode zur Extraktion sind die Histogramm-basierten Verfahren. Sie sind eine konsequente Fortführung der Farbraumtransformation. Basis für das Histogramm ist die Region of Interest für den Mund, die vorab bestimmt werden muss. Die oben beschriebene Gradientenproblematik tritt hier nicht auf. Das Histogramm wird durch einen Schwellwert in Lippenpixel und Nicht-Lippenpixel geteilt (wobei letztere hauptsächlich Hautpixel sind). Der entscheidende Punkt ist die Festlegung des Schwellwertes. Ein sehr einfacher Ansatz, der oft für eine frühzeitige Lokalisierung des Mundes verwendet wird, ist ein fester, nicht-adaptiver Schwellwert, der aus dem Mittel vieler Samples bestimmt wird [7]. Ein etwas flexiblerer Ansatz ist das Watershed-Verfahren. Hierbei wird angenommen, dass der Mund im Histogramm der ROI eine bestimmten Prozentteil der Gesamtmenge einnimmt [9]. Eine weitere Möglichkeit der Histogrammanalyse besteht in der Auswertung der Topologie des Histogramms. So versuchen [5] ein Minimum zwischen zwei Hauptmaxima (welche jeweils für die Menge der Lippen und die der Hautpixel repräsentieren) zu finden, und definieren das gefundene Minimum als optimalen Schwellwert. Neben diesen beiden grundlegenden Methoden gibt es zusätzlich noch hybride Ansätze, die versuchen die Vorteile beider Verfahren zu vereinen [2, 3].

In der vorliegenden Arbeit wird die Mund-ROI als gegeben angenommen. Methoden zur Gesichtsdetektion sowie zur Detektion von Komponenten des Gesichtes (u.a. dem Mund) sind in der Literatur hinlänglich beschrieben [13,14].

2 Statistische Untersuchungen zu Farbe und Histogrammen

Wie in Abschnitt 1 erwähnt, wird der Mundbereich oft einer Farbtransformation unterzogen, um den Mund hervorzuheben. In diesem Abschnitt werden Voruntersuchungen präsentiert, die eine ideale Farbtransformation ermitteln sollen. Innerhalb des Mundbereiches gibt es 3 Klassen von Pixeln. A: Lippenpixel, B: Hautpixel und C: Zahnpixel. Die Klassen B und C können der Einfachheit halber zu Nicht-Lippenpixeln zusammen gefasst werden. Basis der Untersuchungen waren 56 Farbbilder unterschiedlicher Personen mit unterschiedlicher Beleuchtung, von denen jeweils die ROI des Mundes ausgeschnitten wurde. Innerhalb dieser Bilder wurden dann manuell die oben genannten 3 Klassen markiert.

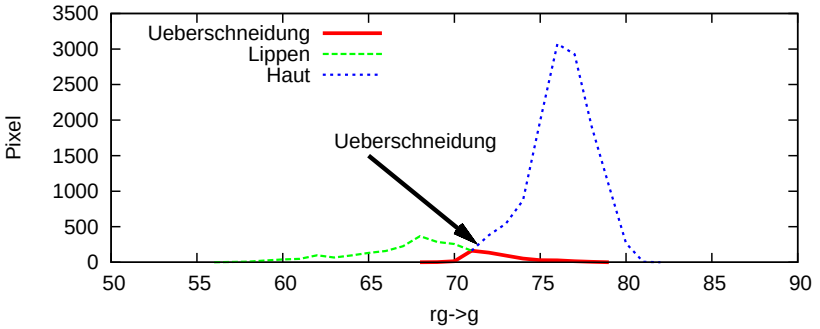
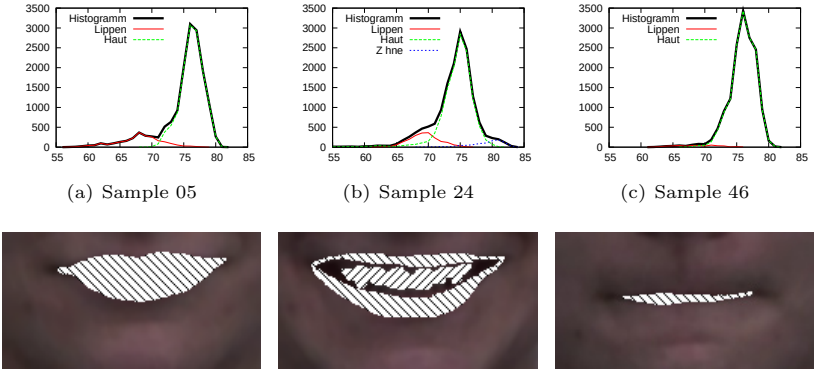


Abbildung 1: Histogram eines Samples. Klasse C (Zähne) ist hier nicht abgebildet. Die Schnittmenge (rot) ist zu minimieren, um eine rein farbbasierte Klassifizierung zu ermöglichen.

Es existiert demnach zu jedem Pixel eine Grundwahrheit, die festlegt, welcher Klasse es angehört. Die durchschnittliche Größe der ROI betrug 160x80 Pixel. Die Aufnahmen enthielten verschiedene Mundstati: Offener Mund mit Zähnen, geschlossener Mund und gepresste Lippen (siehe Abb. 2). Nach vorliegendem Wissensstand existieren keine Untersuchungen, die die Auswirkung der Zähne auf das Histogramm bezüglich der Separierbarkeit von Haut und Lippenpixeln untersuchen.

Eine ideale Farbtransformation ist dann gegeben, wenn die Schnittmenge $I = A \cap (B \cup C)$ minimal ist (siehe Abb. 1). Die in Abschnitt 1 referenzierten Arbeiten propagieren verschiedene Farbraumtransformationen, die als Ideal angesehen werden, um den Mund hervorzuheben. Einige der meistgenannten Farbtransformationen sind in Tabelle 1 aufgelistet. Mit Hilfe der erzeugten Grundwahrheiten wurde ausgewertet, welche Farbraumtransformation die kleinsten Schnittmengen bildet. Das Ergebnis der Untersuchungen zeigte, dass der normalisierte Grünkanal $rg \rightarrow g$ optimal (dicht gefolgt vom $R/(R+G)$) die besten Werte erzielte (siehe Tabelle 1). Die schlechtesten Ergebnisse wurden von den auf dem $YCbCr$ Farbraum basierten Farbtransformationen erzielt.

Ein weiteres Ergebnis der Voruntersuchungen war, dass Hautpixel und Lippenpixel grundsätzlich Gauß-verteilt sind. Unter günstigen Bedingungen, bilden sie zwei nahezu unabhängige, optisch leicht separierbare Gaußglocken. In diesem Fall ist es einfach ein Minimum zwischen beiden Kurven zu bestimmen, welches als Schwellwert dienen könnte. Diese Beobachtung rechtfertigt ansatzweise Verfahren wie in [5], wo genau nach diesem Minimum gesucht wird. Als Allgemeinregel kann dies jedoch nicht gelten. Abhängig von den verschiedenen



(a) Sample 05 (b) Sample 24 (c) Sample 46

Abbildung 2: (Untere Reihe): Grundwahrheit der Datenbank (a) Mund normal, (b) Offener Mund mit Zähnen, (c) Zusammengepresste Lippen. (Obere Reihe): Die zugehörigen Histogramme nach Transformation in $rg \rightarrow g$. Die schwarze Linie ist das reale Histogramm. Die farbigen Linien repräsentieren die einzelnen Klassen der Grundwahrheit.

Stati des Mundes kann die Topologie des Histogramms stark varieren (siehe Abb. 2). Es können anstatt eines primären verschiedene lokale Minima entstehen. In anderen Fällen verschmelzen die beiden Gaußglocken stark miteinander, so dass es sehr schwierig ist überhaupt ein Minima zu lokalisieren. Dieses Verhalten konnte unabhängig von den verschiedenen Farbtransformationen beobachtet werden. Das Auftreten von Zähnen hingegen hatte kaum Einfluss auf die Topologie des Histogramms. Im Falle des $rg \rightarrow g$ entsteht manchmal ein kleiner Cluster im rechten Bereich der Hautpixelglocke. Meistens vermischen sich die Zahnpixel jedoch mit den Hautpixeln.

3 Automatische Schwellwertbestimmung

Die Untersuchungen des vorangegangenen Abschnitt 2 ergaben, dass die $rg \rightarrow g$ Farbtransformation optimal für das Problem der farbbasierten Mundsegmentierung ist. In diesem Abschnitt wird nun beschrieben, wie *vollautomatisch*, *adaptiv* ein guter Schwellwert bestimmt werden kann. Für die gewählte Farbtransformation kann die Annahme als sicher gelten, dass die Lippenpixel einen niedrigen Intensitätswert aufweisen als die Hautpixel. Für den Segmentierungsprozess ist es von Vorteil, falsche Lippenpixel zu vermeiden. D.h. es wird vor-

Tabelle 1: Schnittmengen bei verschiedenen Farbtransformationen

Source	**Transf.**	**With teeth**	**No Teeth**
[2]	$Luv \rightarrow u$	4.61%	3.16%
[11]	G/B	5.97%	2.59%
[3]	G/R	2.30%	1.45%
[4]	Cr^2	11.75 %	10.97%
[4]	Cr/Cb	13.08 %	11.16%
[9]	$R/(R+G)$	2.36 %	1.48%
not found	$rg \rightarrow g$	**0.09%**	**0.38%**

gezogen einige der Lippenpixel nicht erkannt zu haben, anstatt zu viele. Dies begründet sich aus der generellen Topologie des Histogramms. Ein zu hoher Schwellwert birgt die Gefahr, dass auf einen Schlag sehr viele Hautpixel mit zur Lippenpixelmenge zugezählt werden, da die Anzahl der Hautpixel prinzipiell größer ist.

Sowohl Lippenpixel als auch Hautpixel sind Gauß-verteilt. Überlagern sich beide Verteilungen ist eine rein farbbasierte Segmentierung ohnehin Fehler behaftet. Ziel ist es dann diesen Fehler zu minimieren. Betrachtet man beide Gaußglocken, so ist ein idealer Schwellwert am linken Fußpunkt der Haut-Gaußglocke. Dieser schließt (theoretisch) alle Hautpixel aus und stellt sicher, dass vornehmlich wirkliche Lippenpixel gewählt werden. Ziel des nachfolgend vorgeschlagenen Algorithmus ist es, die Verteilung der Hautpixel zu approximieren und daraus auf den Fußpunkt der Gaußglocke und somit den idealen Schwellwert zu schließen. Das σ einer Gauß-Verteilung lässt sich im allgemeinen aus den Stichproben (in dem Fall den Hautpixeln) bestimmen. Diese vermischen sich hier jedoch im Histogram mit den Lippenpixeln. Deswegen wird ein konservativer „First Guess“ benötigt. Dieser sollte so bemessen sein, dass er mit Sicherheit *keine* Lippenpixel enthält, also exakt das Gegenteil von dem, was am Ende erreicht werden soll. Der „First Guess“ ist wie folgt definiert (Siehe auch Abb. 3) :

Sei $h(x)$ der Wert des x'ten Slots des Histograms und h_{max} der Wert des Histogrammslots mit den meisten Pixeln. Des Weiteren sei

$$\epsilon = h_{max}/\alpha \tag{1}$$

Um das σ_h und den Mittelwert μ_h der Hautfarbverteilung zu bestimmen wird zunächst das σ für alle Pixel x bestimmt dessen Slot $h(x)$ im Histogramm folgende Bedingung erfüllt:

$$h\left(x\right) > \epsilon \tag{2}$$

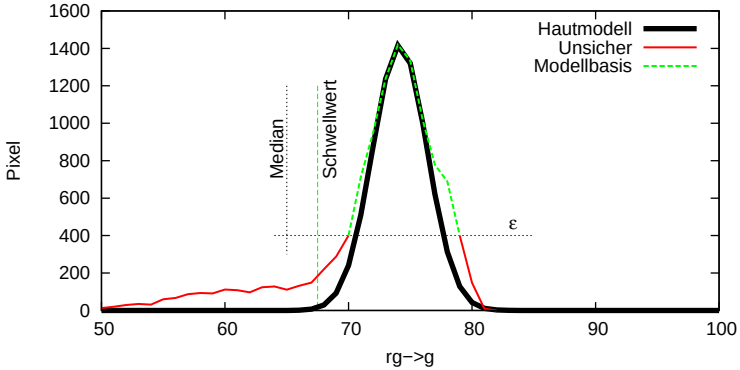


Abbildung 3: Basierend auf den „First Guess“ wird im Histogramm ein Bereich als sicher definiert. Dieser Ausschnitt des Histogramms dient als Basis für die Approximation der Gaußglocke.

$$x > median \tag{3}$$

Die Bedingung in (1) repräsentiert ein erwartetes Verhältnis von Mundgröße zur Größe der ROI wieder. Allerdings sind die Werte so bemessen, dass bei miteinander vermischten Haut- und Lippenpixeln nur mit sehr geringer Wahrscheinlichkeit Lippenpixel mit in Betracht gezogen werden. Ein guter Wert für α ist 3. Es also werden nur Histogramslots akzeptiert, die mindestens mit einem Drittel von h_{max} belegt sind. Um Störungen im Bereich der geringen Intensitäten des Histograms auszuschließen wurde zusätzlich noch eine Median Bedingung eingeführt. Sie soll sicher stellen, dass keine Peaks links des Histogramm-Medians versehentlich in Betracht gezogen werden. Beide Bedingungen in Kombination bieten eine gute Grundlage für einen initialen „First Guess“. Interessant an dieser Stelle ist, dass bereits der sehr einfach zu bestimmende Median einen Schwellwert definiert, der ein respektables Ergebnis liefert (siehe Abschnitt 4).

Die mit dieser Methode bestimmten Pixel bilden die Basis für die Approximation der Hautverteilung. Die Gaußverteilung des so bestimmten σ_g hat wesentlich weniger Streuung, als die wirkliche Hautverteilung. Es besteht jedoch ein Zusammenhang zwischen dem Verhältnis von α in Gleichung 1 und dem Verhältnis von σ_g zum wirklichen σ der Hautpixel. Mit

$$\sigma_h = \sigma_g \cdot \left(1 + \frac{1}{\alpha}\right) \tag{4}$$

erhält man eine gute Approximation der wirklichen Hautverteilung. Je größer

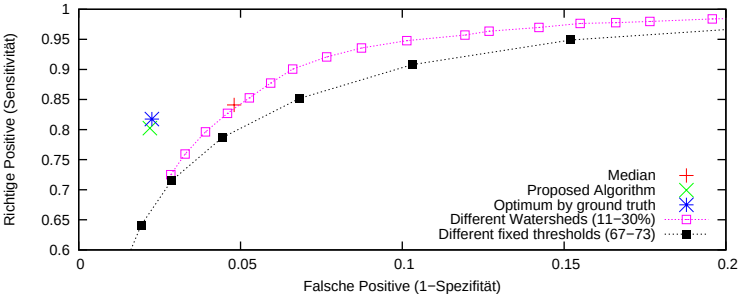


Abbildung 4: Das Kreuz markiert die Qualität des vorgeschlagenen Algorithmus. Die Kurven repräsentieren die ROC-Plots für verschiedene Stufen des Watershed-Algorithmus (10-30%)und für verschiedene feste Schwellwerte.

Tabelle 2: Ergebnisse von Richtige Positiven(RP) und Falschen Positiven(FP) und deren Streuung

Verfahren	RP-μ	RP-σ^2	FP-μ	FP-σ^2
Optimal	81.75%	3.88%	2.262%	0.002%
Vorgeschlagen Adaptiv	80.24%	2.55%	2.200%	0.031%
Watershed 13%	79.16%	1.98%	3.917%	0.223%
Watershed 16%	87.74%	0.93%	5.933%	0.339%

α ist, desto kleiner ist der Teil, der vom Histogramm weg geschnitten wird. Damit steigt natürlich theoretisch die Genauigkeit der Approximation der Gaußverteilung. Geht α gegen Unendlich nimmt man exakt die Grundmenge der Gaußverteilung. Im Anwendungsfall muss jedoch bedacht werden, dass die Lippenpixel heraus gefiltert werden sollen. Die Wahl von α ist demnach eine Abwägung zwischen Approximationsgenauigkeit und dem Wagnis Lippenpixel mit einzubeziehen. Es hängt auch davon ab, wie die Mund-ROI bestimmt wird, und wie das zu erwartende Mengenverhältnis von Haut- zu Lippenpixeln ist. Nachdem die Verteilungsparameter der Hautfarbe σ_h und dessen Mittelwert μ_h bestimmt sind, lässt sich mit der kumulativen Verteilungsfunktion der Fußpunkt der Dichtefunktion bestimmen. Der Schwellwert wurde hier auf 0.01 gesetzt.

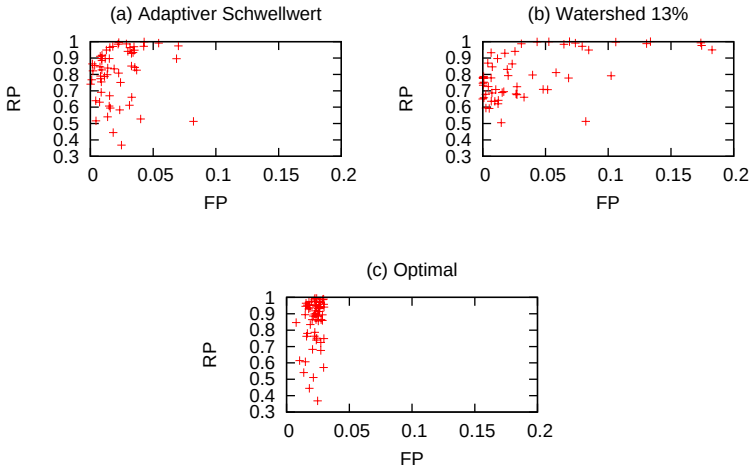


Abbildung 5: Die Verteilung der Einzelergebnisse als im RP/FP Raum. (a) Vorgeschlagener Algorithmus (b). Watershed, (c) Optimaler, manuell bestimmter Schwellwert für jedes Bild mit FPR ($< 5\%$)

4 Experiments and Results

Der in Abschnitt 3 vorgestellte Algorithmus wurde mit zwei ebenso rein farbbasierten Methoden verglichen. Zum einen mit verschiedenen festen Schwellwerten und zum anderen mit einem Watershed Verfahren (ähnlich [5]). Beim Watershedverfahren wurden verschiedene Schwellwerte (10%-30%) getestet. Die festen Schwellwerte schnitten erwartungsgemäß am schlechtesten ab. Das Watershedverfahren konnte besser an verschiedene Bildsituationen adaptieren. Vor allem was unterschiedliche Beleuchtungen angeht. Dennoch bleibt der Nachteil, dass implizit von einem konstanten Verhältnis von Mundgrößte zur Größe der ROI ausgegangen wrid. Am besten schnitt das oben beschrieben Verfahren ab. Um eine ähnlich hohe TP Rate zu erreichen erzeugte der Watershed beispielsweise immernoch doppelt so viele FP wie der vorgeschlagene Algorithmus. Der vorgeschlagen Algorithmus erreicht eine mittlere Trefferquite Richtiger Positiver (RP) von ca. 80%. Das klingt nicht sehr hoch, ist jedoch ein für das gewählte rein farbbasierte Verfahren ein gutes Ergebnis. Als Vergleich dient hier die Hitrate der manuell und für jedes Bild individuell gewählten Schwellwerte. Diese wurden so gewählt, dass maximal 5% FP entstehen und bilden so den jeweils optimalen, bestmöglichen individuellen Schwellwert für jedes Bild. Die Treffer-

quote lag hier auch *nur* bei 81%. Demnach ist der vorgeschlagene Algorithmus sehr nahe an dem, was mit reinen Farbinformationen überhaupt erreichbar scheint (Siehe Tabelle 2).

5 Zusammenfassung und Ausblick

Es wurde eine neue Methode vorgestellt, die zum Zwecke der Mundsegmentierung adaptiv einen Schwellwert bestimmen kann, um innerhalb der Mund-ROI des Gesichtes Haut- von Lippenpixeln zu trennen. Dazu wurde statistisch eine optimale Farbtransformation ermittelt, die der gegebenen Problematik am besten entspricht. Die vorgeschlagene Segmentierungsmethode liefert bessere Ergebnisse als andere rein farbbasierte Methoden, was hauptsächlich auf ihre Adaptivität zurück zu führen ist. Die vorgeschlagene Methode soll keine endgültige Lösung des Mundsegmentierungsproblems darstellen. Vielmehr ist sie als erster Schritt zu werten auf dem Weg zu einer hybriden Methode, die sich sowohl der Farbinformation als auch der Textur und Gradienten bedient. Ziel war es hier den nachweislich maximalen Gewinn aus den Farbinformationen zu ziehen, damit diese in einer zukünftigen hybriden Methode einen maximalen Beitrag leisten können.

Literatur

[1] A. Al-Hamadi, A. Panning, R. Niese, and B. Michaelis. A model-based image analysis method for extraction and tracking of facial features in video sequence. In *The 4th International Multi-conference on Computer Science and Information Technology CSIT 2006, Spo. by IEEE, Aman,Vol.3*, pages 499–509, 2006.

[2] S. Arca, P. Campadelli, and R. Lanzarotti. A face recognition system based on local feature analysis. In *Audio- and Video-Based Biometric Person Authentication*, pages 182–189, 2003.

[3] C. Bouvier, P.Y. Coulon, and X. Maldague. Unsupervised lips segmentation based on roi optimisation and parametric model. In *IEEE International Conference on Image Processing*, pages IV: 301–304, 2007.

[4] Jingying Chen, Bernard Tiddeman, and Gang Zhao. *Advances in Visual Computing*, volume 5359/2008 of *Lecture Notes in Computer Science*, chapter Real-Time Lip Contour Extraction and Tracking Using an Improved Active Contour Model, pages 236–245. Springer Berlin / Heidelberg, 2008.

[5] P. Cisar and Zelezny M. Using of lip-reading for speech recognition in noisy environments. In *Speech Processing*, pages 137–142, Prague, 2004. Academy of Sciences of the Czech Republic.

[6] N. Eveno, A. Caplier, and P.Y. Coulon. Accurate and quasi-automatic lip tracking. *Circuits and Systems for Video Technology*, 14(5):706–715, May 2004.

[7] Erhan AliRiza Ince and Syed Amjad Ali. An adept segmentation algorithm and its application to the extraction of local regions containing fiducial points. In *ISCIS*, pages 553–562, 2006.

[8] K.S. Jang, S. Han, I. Lee, and Y.W. Woo. Lip localization based on active shape model and gaussian mixture model. In *Pacific-Rim Symposium on Image and Video Technology*, pages 1049–1058, Hsinchu , TAIWAN, 2006.

[9] J.Y. Kim, S.Y. Na, and R. Cole. Lip detection using confidence-based adaptive thresholding. In *International Symposium on Visual Computing*, pages I: 731–740, 2006.

[10] S.H. Leung, S.L. Wang, and W.H. Lau. Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Transaction on Image Processing*, 13(1):51–62, January 2004.

[11] Trent W. Lewis and David M.W. Powers. Lip feature extraction using red exclusion. In Peter Eades and Jesse Jin, editors, *Selected papers from Pan-Sydney Area Workshop on Visual Information Processing (VIP2000)*, volume 2 of *CRPIT*, pages 61–67, Sydney, Australia, 2001. ACS.

[12] D. Nguyen, D. Halupka, P. Aarabi, and A. Sheikholeslami. Real-time face detection and lip feature extraction using field-programmable gate arrays. *IEEE Trans. Systems, Man and Cybernetics, SMC-B*, 36(4):902–912, August 2006.

[13] A. Panning, A. Al-Hamadi, R. Niese, and B. Michaelis. Facial expression recognition based on haar-like feature detection. *Pattern Recognition and Image Analysis*, 18(3):447–452, 2008.

[14] Paul Viola and Michael Jones. Robust real-time object detection. *Second international workshop on statistical and computational theories of vision - modeling, learning, computing and sampling*, 2001.