REAL-TIME LIPTRACKING FOR SYNTHETIC FACE ANIMATION WITH FEEDBACK LOOP

Franck Luthon
University of Pau and Adour River, Computer Science Lab. LIUPPA
IUT Château Neuf, Place Paul Bert, 64100 Bayonne, France
http://www.iutbayonne.univ-pau.fr/∼luthon/
Franck.Luthon@univ-pau.fr

Brice Beaumesnil
University of Pau and Adour River, Computer Science Lab. LIUPPA
IUT Château Neuf, Place Paul Bert, 64100 Bayonne, France
beaumesn@iutbayonne.univ-pau.fr

Keywords: Segmentation, Closed-Loop, Hue, Motion, Snake, Active Contour, Talking Head, 3D-Model.

Abstract: This article deals with facial segmentation and liptracking with feedback control for real-time animation of a synthetic 3D face model. Straightforward approaches consist in two successive steps: video analysis then synthesis. Our approach departs from the previous ones in that we build a global analysis/synthesis processing loop, where the image analysis needs the 3D synthesis and conversely. A first facial segmentation is computed according to which the 3D face model is positionned. Then the feedback loop, implemented from the 3D animated model back to the input pixel segmentation algorithm, helps to correct some (few) control points that were badly tracked, which are detected by measuring the vertical distance between lip contour and corresponding 3D face model. When this distance is too big, we re-enter into the image segmentation process and zoom-in inside a few regions of interest (ROI) where the algorithm is run again, with a new set of tuning parameters better suited to the pixel neighborhood context. In that way, the face segmentation is refined in order to extract more precise parameters. This approach is inspired from control theory with closed-loop systems. The contribution of the paper is to use simple image processing techniques, but to improve segmentation through the feedback loop. Results show that real-time and robust performances are achievable under real-world conditions, which are two key issues for face and lip tracking applications.

1 INTRODUCTION

We present a complete real-time analysis/synthesis framework allowing lip tracking for animation of a clone with a single camera in unconstrained environment (typ. webcam in the office). The approach is based on lip segmentation from a hue component computed within a non-linear color space that is robust to luminosity variations. Inner and outer active contours are extracted, and then interpreted to make a real-time realistic animation of a clone's mouth.

For synthetic talking head animation, realistic rendering of lip motion is the key point. The purpose of this paper is to link in real-time the face image analysis with the synthesis of an animated 3D model of the head. We focus on the speaker's lip video segmentation from a mono-camera (webcam or motorized camera) for on-line animation of the mouth of a clone, without dealing here with the sound information (no speech processing). In future developments of course, speech processing should be coupled with image processing for optimal audiovisual rendering. But here, we want to investigate what one can do with the video only. Our aim is to get a realistic rendering of the mouth motion, but not necessarily to get the most precise lip contour extraction. It means that image analysis needs not to be very sophisticated, but just **good enough** for our application (*i.e.*, get an acceptable rendering for a realistic animation). We propose to compensate for some defaults in the analysis by implementing a feedback loop form the 3D synthesis towards the input segmentation process. In other words, the low-level (early vision) segmentation process is corrected (on-line at video rate) by some high-level information coming back from the 3D synthetic face (with all its semantics and animation constraints) through a retroaction channel. Hence, the key point of our scheme is not to use or develop complex and time consuming algorithms, but to use a relatively simple segmentation algorithm, and to take advantage of the feedback channel to improve the segmentation quality in areas where it failed at the first run. Therefore, our efforts are dedicated to the design of the global real-time processing chain (from video input towards synthetic clone output).

Most of the approaches proposed in the literature

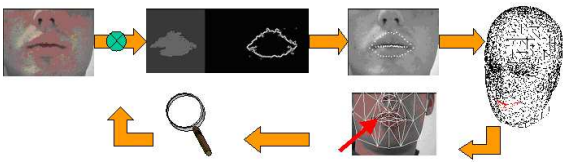

Figure 1: Feedback scheme for real-time coupling of liptracking with animation

do not implement this feedback loop. They are often based on high level methods like AAM (Cootes et al., 1998), that require large-sized video training databases for off-line learning. If they are well suited for the whole face (Batur and Hayes, 2005), they are not precise enough for the lips.

Our approach departs from the above mentioned methods in the sense that we want to use low-level algorithms that can adapt to any mouth shape without a priori knowledge about the face. The main adavantages of such a strategy is that real-time is achievable and that neither (possibly heavy) learning stages nor huge databases (to derive a priori knowledge about faces) are required.

In the literature, some authors have already proposed to use a feedback strategy, either for texture segmentation (Pichler et al., 1998) (with K-means classification in the open-loop, smoothing filter in the closed-loop), object recognition (Mirmehdi et al., 1999) (with three control strategies at low, intermediate and high level). For video object tracking (Erdem et al., 2003) proposes a scheme with boundary prediction in the open-loop, boundary correction in the closed-loop, and performance measures without ground-truth (Erdem et al., 2004) that yields some nice properties and results: non rigid object tracking, robustness to occlusions, no need of training thanks to automatic weight control, on-line tracking for coarse estimates. However, pixel accurate boundary tracking is only achieved off-line. Moreover, the use of active contours requires a proper initialisation (via user interface as in (Fu et al., 2000)). In (Erdem et al., 2003), the boundary initialisation on the first frame is done manually by the user in an interactive (non-real time) mode. In real-time mode, a change detection algorithm is proposed but not implemented in real-time.

Taking experience from those previous works, we propose a contribution in liptracking to remove the constraint of manual initialisation or training and to achieve real-time processing with pixel-accuracy.

2 Description of the Processing Loop

Fig. 1 illustrates our global approach. The forward stage of our framework is divided into four parts:

1. low-level color segmentation (analysis): it works in a color space that is little sensitive to lighting variations and exhibits very distinctly skin and lip hue areas.
2. active contour positioning (estimation): to delineate both inner and outer lip contours.
3. transmission of geometrical parameters (communication): via Internet to a distant computer
4. clone animation (interpretation/synthesis): it uses the lip contours extracted at step 2 and transmitted at step 3.

The backward stage consists in three steps:

5. error measurement at the receptor side
6. backward transmission of some ROIs that need to be processed again (because of poor segmentation)
7. re-run of the segmentation process on those ROIs (step 1)

Our work assumes that we have at our disposal an automatic tool for face detection and tracking, so that an optimal framing of the speaker's face is available (which is required in the case of videoconferencing for example). This is part of another project about intelligent videoconferencing that is under study at our laboratory[1]. If this is not the case, the only constraint is that the speaker should stay in front of the camera with little motion (normal behavior in front of a webcam). In this paper, we simply ask the speaker to seat in front of the camera so that his face covers the major part of the image. Moreover, we ask him to have the mouth closed (neutral position) on the first frames of the sequence, in order to make a good initialisation of his mouth proportions (width and height). This allows to take into account the distance between the speaker and the camera, that may vary from one acquisition to another, but that is important for 3D model calibration and clone animation. Therefore, the whole image frame is considered as the search area at the beginning.

Then the speaker should stay at about the same distance from the camera during the session, so that the mouth shape proportions remain the same (since the camera is supposed to be static here).

Lighting conditions are not calibrated nor constant: they correspond to realisitic office environment (non uniform lighting, light sources that may be added or suppressed depending on the daylight).

Starting from two previous works: an MRF classification approach based on hue and motion detection (Liévin and Luthon, 2004), and an active contour approach by Delmas *et al.* (Liévin et al., 1999), we have developed our framework, that connects synthesis with anlaysis through a feedback loop, in order to

[1] Laboratoire d'Informatique de l'Université de Pau et Pays de l'Adour, France, http://liuppa.univ-pau.fr/

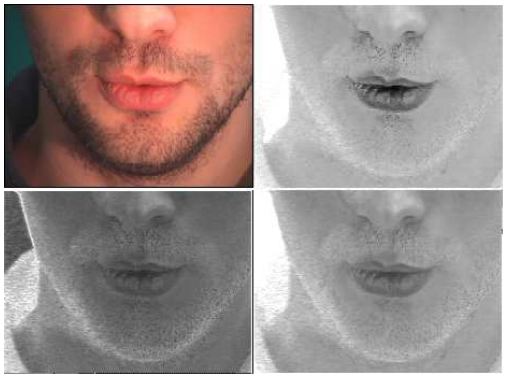

Figure 2: a) Original image; b) Hue computed from LUX; c) Hue from HSL color space; d) Hue computed as G/R (often used for face/lip segmentation)

counteract the defaults of the image analysis part and achieve the following goals:

- good performance in unconstrained environment (ordinary lighting conditions, small head-movements allowed)
- real-time running of the analysis/synthesis loop (*i.e.* at video rate: 25 or 30 fps)
- enough precision on the inner lip contour for realistic animation
- to cope with hair, beard, race, gender.

2.1 Extraction of Lip Area

For face and lip segmentation, we work in the LUX color space (Liévin and Luthon, 2004). This color space is non-linear with respect to the RGB color components. It helps to reinforce the color contrast while being relatively insensitive to lighting variations. In this color space, most of the color information relative to a human face is coded by the hue component (red chromaticity) in the particular case when $R > L$ (L being the luminance Eq. 1). Therefore, we derive from LUX space the simplified hue component U, Eq.(2), that is more discriminating than RGB or HSL for face and lip segmentation (see Fig. 2).

$$L = (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1 \quad (1)$$

$$U = \begin{cases} 256\frac{L+1}{R+1} & \text{si } R > L, \\ 255 & \text{otherwise or if } L < \lambda\ . \end{cases} \quad (2)$$

The threshold λ is introduced in order to detect very dark areas like the inner side of the mouth or nostrils. This will amplify the various gradients computed on the hue, as explained below.

Since the hue difference between face and lip is well contrasted in LUX space, we can easily classify pixels as lips or face by simple use of the K-means classification algorithm. The algorithm works with three classes : lips, face and background; it exploits two types of low-level information: the mean value of hue in a given neighborhood, and the maximum deviation from this mean value for any pixel in the neighborhood. Compared to MRF-based classification that induces high computing time, the choice of the K-means technique is justified be its lower computation cost.

But a good initialisation of the centers of the three classes is required for proper convergence of the algorithm. Since we want to avoid heavy learning stage with huge data-bases, we use an heuristic approach to set ad hoc parameters that only depend on the mean-value of speaker's face hue in the whole image (that is easily estimated through histogram computation). This allows a good initialisation of the class centers, and hence a rapid convergence towards the solution (Fig. 3d).

In addition, we have implemented a non linear filter that searches for the lip bounding box (BB) in the following way: since the mouth is the bigger reddish horizontal form in the face, we look for this kind of feature inside the face among pixels belonging to the lip-class. We scan sequentially all pixels belonging to the lip-class and apply the nonlinear filter mask defined by Eq.(3). If the neighbors belong to the same class (lip), their filtered value is weighted according to their position and contributes to the current pixel if it also belongs to lips. Otherwise, the current pixel is set to zero.

$$M(i) = \begin{cases} 1 + \sum_{j \in \nu(i)} a_j M(j) & \text{if } U(i) \in \text{lip} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

M is a matrix of size $L \times C$ (image size) initialised to 0 and the a_j (with $j \in \nu(i)$) are the coefficients attributed to the different pixels belonging to the causal and connex neighborhood $\nu(i)$ of the current pixel i (see Fig.3a).

This nonlinear filter technique yields in a single pass (rapid processing) the localisation of the lips. The mask gives more importance to elongated forms in the horizontal direction, which allows to make the difference between ears and mouth, that are usually the two biggest reddish zones in a face (Fig. 3d). After filtering, we get a map of the face with the higher values located on the largest reddish forms, among which the mouth with the highest value of the map (usually at the lower left edge of the lips, see white cross in Fig. 3d).

This simple technique gives an approximate BB for mouth positionning in the image plane (Fig. 3b). However, it may fail when lip corners are not well detected or when lower and upper lips form two separated areas (this may happen e.g. in the case of the French phoneme [a]). To cope with those situations,

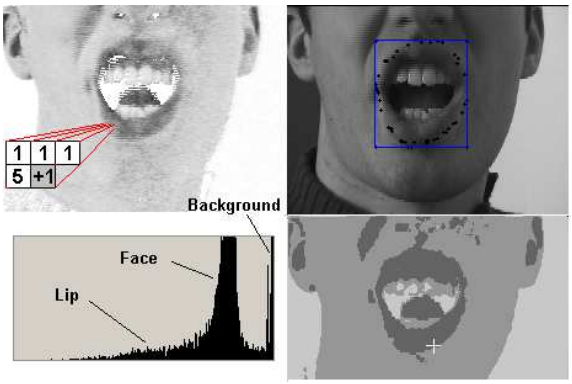

Figure 3: a) Hue U with non-linear filter mask coefficients a_j ; b) Original color image, with bounding-box and outer lip contour points; c) Hue Histogram; d) 3-class map given by the k-means algorithm, with pixel having the highest value of M in white

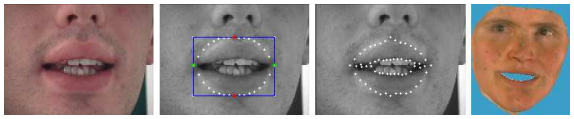

Figure 4: Liptracking for real-time animation of a 3D face model: a) Video input: color image ; b) Initialisation of outer snake obtained from bounding box (in blue); c) Inner and outer snakes after convergence; d) Animated clone: synthetic talking head position

a lip tracker is added: it is based on Lucas-Kanade algorithm applied on a few relevant points detected on the outer lip contour. Since the speaker is supposed to have his mouth closed in the first image, we are guaranteed to have a single lip area at the beginning, so that the outer lip contours can be tracked properly afterwards, even if the two lips are no more connected areas in the video sequence. Using this simple motion estimation technique instead of motion detection as in (Liévin and Luthon, 2004) yields a better robustness to lighting variations.

Having this complementary information, we are able to complete the face classification map and get the whole mouth area properly: we look for one or two forms close to the tracked points and we link them together. We then get the mouth optimal BB by a simple scan of the connected component lip-hue area that includes the pixel detected as being the most likely on the lip border (point giving the maximum value of M, *i.e.* located on the bigger horizontal reddish form, Fig. 3d).

2.2 Snakes and Clone Animation

For estimating the lips outer border, one active contour or snake (Kass et al., 1987) is initialised with the BB detected as explained above.

This snake is made of a finite number of control points (typ. $2 \times 16 = 32$ points) that are forced to undergo only vertical displacements during iterations. The points are initialised on cubic curves computed from the BB and from the lip map (to position the lip corners), Fig. 4b. The forces used for snake convergence are the following:

- an internal force that controls elasticity and curvature (very classicaly defined)
- an external force that specifies the features that should attract the snake (namely spatial gradients computed both on hue and on luminance maps)
- a constraint force that is specific of the problem at hand (the snake is forced to converge towards the gravity center of the BB).

After convergence of the outer snake, another snake (inner one) is initialised on the outer one, then shrinked by an anisotropic scaling w.r.t. the mouth center and taking into account the actual thickness of lips.

After convergence of the inner snake towards the inner lip contour, the interpretation step for our application purpose is readily done: we are able to compute various geometrical features from the control points of the two snakes, that are then used as input parameters for the 3D head model. This computation of animation parameters is of course dependent on the 3D model used. Our talking head is a 3D clone from ICP[2]. It is built with 275 mobile points that allow realistic synthesis of mouth motion thanks to six animation parameters (dedicated to visemes and phonemes of the French language) (Benoît et al., 1992). In our case, a simple system of linear equations transforms the snakes control points coordinates into animation parameters that are taken as input by the 3D model. Actually, we only use four control points: left and right lip corners, central upper and lower lip positions.

3 Experimental Results

We present in this section two types of results:

- image analysis results that show the performance of lip segmentation, BB detection, lip contours estimation (Fig. 5)
- image synthesis results that show various positions of the animated model computed/interpreted from the points estimated in the first step, and show also some cases of failure that are corrected by the feedback loop.

[2]Thanks to Institut de la Communication Parlée, Grenoble, France, http://www.icp.inpg.fr

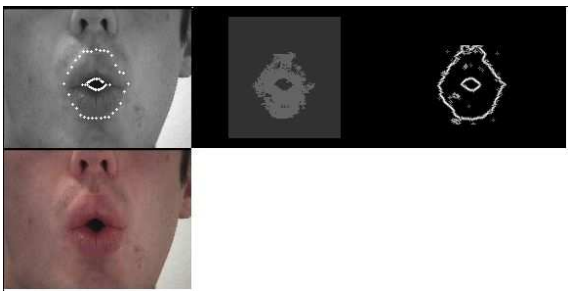

Figure 5: a) Snakes after convergence; b) lip hue classification map; c) gradients; d) color image

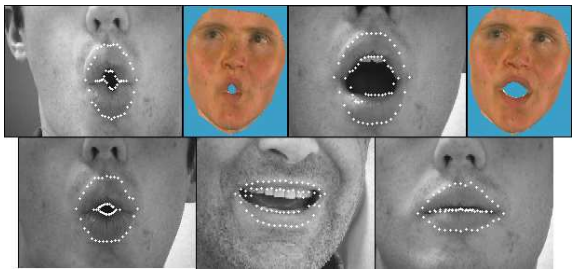

Figure 6: Typical analysis/synthesis results for different visemes corresponding to French phonemes: top) vowels $[o]$ and $[a]$ with corresponding synthetic clone positions; bottom) vowels $[e]$ and $[\partial]$ plus the neutral position

3.1 Analysis Result: Segmentation

Fig. 2 shows the interest of using color components derived from the non linear LUX color space, especially for skin hue segmentation in adverse lighting conditions: one can see that lip hue is clearly distinct from the face skin hue (see Fig. 2b), even in the case of a non uniform lighting (here the face is enlighted sideway).

The algorithm is robust to lighting variations: abrupt light changes (like putting on/off a lamp in the room) do little influence the quality of segmentation since relevant hue mean values (for each of the three classes) are re-estimated on-line as each new frame is acquired, so that the classification algorithm adapts to light changes over time.

Since our hue-based algorithm does not rely on any a priori knowledge about the face (apart from the mean value of hue that is estimated online), it learns special features like beard or mustaches at initialisation (a case of hairy face is illustrated in Fig. 6).

This lip segmentation process yields precise location of the mouth in the face, with a BB including completely the lips, which gives a good initialisation for the snakes.

3.2 Synthesis Result: Animation

After BB detection and snake convergence on the lip contours (Fig. 4 and 5), one can interpret the detected points for lip animation of the 3D model.

The keypoint is that even if lip contours are not very well estimated (poor initial segmentation), it is not redhibitory for a realistic animation. Indeed the comparison between the estimated snake locations and the 3D model positions allows to make a backward correction (through the feedback loop).

As regards the inner lip contour, teeth are not a problem: they even amplify the gradients, ensuring a better convergence (see Fig. 6). However, another issue that has to be addressed is the presence of reflected highlights on lips (Fig. 6). The proposed feedback processing helps to solve this problem (Fig. 7).

As a matter of fact, outer lip contours are very well estimated when the mouth is wide open. On the opposite, inner lip contours are well estimated when the mouth is closed or in the case of protrusion We may use this fact in conjunction with the 3D result to guide the processing of re-computation where the segmentation initialy failed. This is what we are currently investigating.

4 Discussion and Future Work

We have shown that real-time realistic animation of a synthetic talking head is achievable with :

- a rapid initialisation without heavy learning stage,
- a few points tracked by a fast segmentation scheme
- the use of a feedback loop to correct some segmentation defaults.

Currently, only vowels were tested carefully. The clone reproduces the speaker's lip motion with a very nice precision. It moves even to fast ; we have to slow down the animation process in order to reproduce realistic human lip motion: in fact, this will be very easy to implement because of the feedback control. As in the case of control systems (cf. PID correctors in automation theory), one can give to the global chain the right time constant and damping factor in order to guarantee stability and to reach the best compromise between speed and precision.

This work demonstrates that one can build a complete anlaysis-synthesis chain that works in real-time for 3D head animation, without using sound information nor face databases for learning.

The algorithmic part is made of relatively simple (and hence rapid) image processing steps: LUX color transform for robustness to lighting conditions and lip separation, K-means classification with 3 hue classes, non linear filter for BB initial localisation,

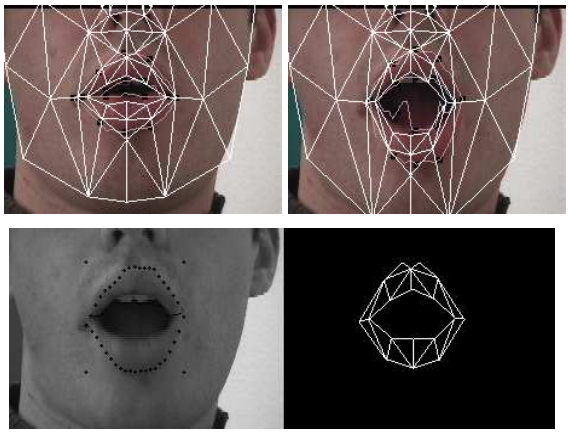

Figure 7: Regularisation through the 3D model: a) Good match between segmentation and model; b) Bad match: a control point of the inner snake is clearly erroneous and is corrected by the 3D mesh corresponding position ; c)d) After correction through feedback: front view with 3D lip model reshaped

motion estimation via LK algorithm for Bounding Box tracking, snake initialisation deduced from the BB position, snake convergence towards inner and outer lip contours. The (eventually) badly segmented points (detected by a distance measure on the 3D model) are re-processed (thanks to the closed-loop). Currently, we simply replace the bad point by the corresponding point of the model. But we will now implement a reprocessing on that pixel (zoom-in in the neighborhood and local re-computation in that little area).

Indeed we still have some time left to spend at this reprocessing: the whole analysis algorithm, implemented in non optimised C-code on an $i386$ processor at $1.4GHz$, works in real-time (i.e. processing rate better than $30Hz$). This is 30 times faster than the algorithm presented in (Liévin and Luthon, 2004).

Another direction of our current research is to segment not only the lips, but also other face features (namely nostrils, eyes, eyebrows and ears). This will help to get other relevant points for 3D model scaling (cf. user-dependent facial geometry taken into account at initialisation step), for regularisation of poor input data (cf. re-segmentation through the feedback loop that adds rigidity constraints and other semantics to the ill-posed problem at the pixel side) and for more realistic animation (cf. face expressions during speech).

Moreover, having more information on the whole face may enable a better understanding of some spoken phonemes (e.g. to make the difference between French phonemes [e] and [y], the nose position is very important for animation: indeed, from a visual perception viewpoint, the mouth has almost the same shape in both cases. The only difference is that for phoneme [y], the mouth is closer to the nose

Finally, as we want to be able to animate various clones and propose a generic solution, we are also working on the use of MPEG4-compliant 3D models (using FDP and FAP, facial animation parameters).

REFERENCES

Batur, A. U. and Hayes, M. H. (2005). Adaptive active appearance models. *IEEE Trans. on Image Processing*, 14(11):1707–1721.

Benoît, C., Lallouache, T., Mohamadi, T., and Abry, C. (1992). A set of French visemes for visual speech synthesis. In Bailly, G., Benoît, C., and Sawallis, T., editors, *Talking Machines: Theories, Models and Designs*, pages 485–504, Amsterdam, North-Holland. Elsevier Science Publishers B.V.

Cootes, T., Edwards, G., and Taylor, C. (1998). Active appearance models. In *European Conference on Computer Vision, ECCV*, volume 2, pages 484–498. Springer-Verlag.

Erdem, C. E., Sankur, B., and Tekalp, A. M. (2004). Performance measures for video object segmentation and tracking. *IEEE Trans. on Image Processing*, 13(7):937–951.

Erdem, C. E., Tekalp, A. M., and Sankur, B. (2003). Video object tracking with feedback of performance measures. *IEEE Trans. on Circuits and Systems for Video Technology*, 13(4):310–324.

Fu, Y., Erdem, A. T., and Tekalp, A. M. (2000). Tracking visible boundary of objects using occlusion adaptive motion snake. *IEEE Trans. on Image Processing*, 9(12):2051–2060.

Kass, M., Witkin, A., and Terzopoulos, D. (1987). Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331.

Liévin, M., Delmas, P., Coulon, P. Y., Luthon, F., and Fristot, V. (1999). Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme. In *IEEE Int. Conf. on Multimedia Computing and Systems (ICMCS'99)*, pages 691–696, Firenze, Italy. Vol. 1.

Liévin, M. and Luthon, F. (2004). Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video. *IEEE Transactions on Image Processing*, 13(1):63–71.

Mirmehdi, M., Palmer, P. L., Kitler, J., and Dabis, H. (1999). Feedback control strategies for object recognition. *IEEE Trans. on Image Processing*, 8(8):1084–1101.

Pichler, O., Teuner, A., and Hosticka, B. J. (1998). An unsupervised texture segmentation algorithm with feature space reduction and knowledge feedback. *IEEE Trans. on Image Processing*, 7(1):53–61.