International Workshop on
Coding Techniques for Very Low Bit–rate Video (VLBV 94)
Colchester, United Kingdom, April 1994

AUTOMATIC ADAPTATION OF A FACIAL MASK IN AN ANALYSIS–SYNTHESIS CODER BASED ON MOVING FLEXIBLE 3D OBJECTS

Jörn Ostermann, Markus Kampmann
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover, Appelstraße 9A, 30167 Hannover, F.R.Germany
Phone: ++49–511–7625308, email: osterman@tnt.uni–hannover.de

Abstract

For object–based analysis–synthesis coding (OBASC) [1][2], an image analysis algorithm is required which automatically generates the parameter sets motion, shape and colour needed for synthesizing an image sequence. This contribution assumes 3D objects arbitrarily moving in 3D space. Thus, the 3D real world is modelled by a 3D model world and moving real objects are modelled by moving model objects. It is the goal of the modelling to generate a model world which has a model image identical to the real image. In the past, the coder used the source model flexible 3D objects moving in the real world. Flexible objects are modelled using a 3D wireframe and consist of flexibly–connected rigid components which have a flexible surface [3][4]. No prior knowledge of the object types was assumed. In this contribution, the coder is extended by a knowledge–based mode. Furthermore, an algorithm for detecting human faces in an image sequence is presented. As soon as this algorithm detects moving eyes and mouth of a face in an image, the generic facial mask ”Candide” [8] is adapted to the estimated eye and mouth positions of the face. The facial mask is then incorporated into the corresponding model object. This allows the coder to switch from its object–based mode to its knowledge–based mode [5][6]. In the knowledge–based mode, the coder uses the source model of flexible 3D objects with the facial mask ”Candide”. Inside and outside of the facial area, the coder applies different criteria for evaluating subjective image quality in order to reduce the overall bit–rate or to improve subjective image quality when compared to the object–based mode.

1 The Facial Mask ”Candide”

In this contribution, the human face is modelled by a 3D facial mask derived from ”Candide” [8]. ”Candide” was chosen because it only requires the positions of eyes and mouth to be adapted to a face. Other facial masks, which are better suited for image synthesis, require to fit up to 37 points to a face which is an extremely difficult task for automatic image analysis [9]. Once the mask is adapted to the face, it allows to model roughly the mimic of a person. The shortcomings of ”Candide” as far as image synthesis is concerned, can be overcome by image pattern prototypes [10][11].

2 Face Detection

In order to use the facial mask, it must be adapted automatically to the human face in an image sequence with high precision. The algorithm for face detection is based on [7]. The two main assumptions are:

1. A speaking person moves his head and in addition the eyes and mouth.
2. A real face has two eyes located symmetrically above one mouth.

According to these assumptions, the following steps are carried out:

2.1 Detection of Areas with Moving Facial Features

According to assumption 1, the eyes and mouth are located inside the silhouette of a moving object. Furthermore, it is assumed that there are horizontal contours close to the potential eye and mouth centers. Since the mimic of eyes and mouth cannot be modelled with the source model of moving flexible 3D objects (F3D), the detection of areas with moving facial features computes the difference signal between the original image s_k and the motion–compensated and shape–updated model image $\hat{s}_k$ (prediction image) of the model world. Usually, there are several separate areas with high vertical gradients. In order to cut down the number of areas, only those are accepted whose centers roughly combine to an isosceles triangle (assumption 2).

In order to be able to detect eye and mouth centers even though the mimic does not change from frame s_{k-1} to s_k, the areas with high vertical gradients detected at time instant k–1 are motion–compensated using the motion parameters of the model object. Then, these areas are added to the areas detected at time instant k. The subsequent algorithms for the detection of eyes and mouth are applied to these areas which might contain the facial features eye and mouth.

2.2 Detection of Potential Eye Positions

Within the areas with facial features, a probability measure f_{eye} is evaluated:

$$f_{eye}(x,y) = w_1 \cdot \frac{255 - s_k(x,y)}{255} + w_2 \cdot c^k_{eye}(x,y) + w_3 \cdot C^k_{eye}(x,y) \tag{2.1}$$

The first term assigns a high value to dark pels (x,y) of the real image s_k. The correlation $c^k_{eye}(x,y)$ between the real image signal and an artificial eye template is computed in a window centered at (x,y). Hence, f_{eye} increases with increasing correlation. The third term $C^k_{eye}(x,y)$ is the average of all motion–compensated correlations $c^k_{eye}(x',y')$ according to

$$C^k_{eye}(x,y) = \frac{1}{k-1} \sum_{i=1}^{k-1} c^i_{eye}(x',y') \tag{2.2}$$

where (x',y') are the appropriately motion–compensated coordinates of (x,y). The weighting factors w of Eq. (2.1) have been empirically selected to w_1=3, w_2=2 and w_3=1. Those two points with the highest values of f_{eye} are selected as potential eye positions.

2.3 Detection of a Potential Mouth Position

For the detection of the mouth, it is assumed that there is a high correlation between the image signal and an artificial mouth template. A probability measure f_{mouth} is evaluated:

$$f_{mouth}(x,y) = c^k_{mouth}(x,y) \tag{2.3}$$

where $c^k_{mouth}(x,y)$ is computed similar to $c^k_{eye}(x,y)$. The point with the highest probability measure f_{mouth} is selected as candidate for the mouth center position. The probability measure f_{mouth} does not evaluate a temporal memory because the correct mouth position is more easy to detect than the eye positions.

2.4 Verification of Eye and Mouth Positions

The image coordinates for the potential eye and mouth positions must span an isosceles triangle. Furthermore, some anatomic relations between eye and mouth positions must be fulfilled. For example, the vertical distance between the eyes and the mouth has to be between one and two times the eye distance. This assures that the face is not looking to the side but approximately into the camera.

The evaluation of a single image may not provide enough information for a reliable estimation of the eye and mouth positions. If the verification test fails, the algorithms are applied to the next image. Since the results of the correlation measure $c^k_{eye}(x, y)$ are integrated over the image sequence according to Eq. (2.2), the reliability of the results increases with time.

2.5 Adaptation of the Facial Mask

As soon as the verification test is successful, i.e. the image coordinates of eyes and mouth are known, the facial mask "Candide" is scaled horizontally and vertically in order to adapt it to the face. Then, the mask is positioned on the corresponding model object in the 3D model world. Finally, the mask is incorporated into the 3D model object which was used for coding the moving object prior to face detection.

3 Extension of Object–Based Analysis–Synthesis Coding to Knowledge–Based Analysis–Synthesis Coding

In the object–based mode, the coder applies the source model of moving flexible 3D objects. Based on this source model, image analysis detects moving objects in the scene and describes each object by means of shape, motion and colour parameters. Objects whose correct motion and shape parameters can be estimated are denoted as *model compliance* (*MC*) *objects*. After estimation of motion and shape parameters, image analysis compares the real image s_k to the prediction image $\hat{s}_k$, synthesized using the previously transmitted colour parameters and the new shape and motion parameters. Due to the object–based image description, this prediction image $\hat{s}_k$ gives a subjectively pleasant impression. Image areas of s_k which cannot be described by the MC–objects are detected. These areas of *model failures* (MF) are defined by 2D–shape and colour parameters only and are referred to as *MF–objects*. The detection of MF–objects exploits that small position and shape errors of the model objects – referred to as *geometrical distortions* – do not disturb subjective image quality. Thus, MF–objects are limited to those image areas with significant differences between the prediction image $\hat{s}_k$ and the current real image s_k. They tend to be small in size. This allows to code colour parameters of MF–objects with high quality, thus avoiding subjectively annoying quantization errors. Since the transmission of colour parameters is expensive in terms of data rate, the total area of MF–objects should not be more than 4% of the image area at 64 kbit/s, CIF and 10Hz. The average size of model failures determines the quality of an analysis–synthesis coder since coding of colour parameters requires at least 1.3 bit/pel for acceptable image quality.

As soon as the face is located, OBASC can switch to an knowledge–based mode. In this mode, it is no longer necessary, to apply one unique criterion for detecting model failures to the whole image [12]. Now, it can be distinguished between subjectively important image areas like the face and less important areas. Hence, it is possible to concentrate the model failures to the facial area and thereby to decrease the average areas of model failures allowing for larger position and shape errors outside the facial area. This will not give a decrease in subjective image quality because the object–based source model assures that a prediction image gives a subjectively correct impression.

4 Experimental Results

A video tape will be shown demonstrating the performance of the face detection. For the test sequences *Miss America* and *Claire* at CIF, 10 Hz, it takes app. 10 frames to detect the complete silhouette of the moving objects. After 2 to 5 more frames, the facial mask is adapted to the image sequence. After the coder switches to the knowledge–based mode, the average area of model failures decreases from 3% to 2.5% of the image area. Coding of model failures requires the transmission of colour parameters which takes most of the available bit–rate. Thus, decreasing the average area of model failures gives significant advantages for coding of videophone image sequences.

5 Conclusion

An algorithm for the detection of the eye and mouth position of human faces has been presented. As soon as the face is localized, the facial mask "Candide" is adapted to the image sequence allowing the coder to switch from an object–based mode to a knowledge–based mode. For videophone image sequences showing a human speaker, exploiting the knowledge about the face position reduces the average area of model failures, for which an update of the colour parameters has to be transmitted, from 3% to 2.5%. Since the coder switches to the knowledge–based mode only if a face is detected, the existence or detection of faces is not a prerequisite for the coding algorithm to work properly.

In the future, research will be focussed on the modelling of the human face and the estimation of mimic parameters.

This work has been supported by the Deutsche Bundepost Telekom.

6 Literature

[1] H.G. Musmann, M. Hötter, J. Ostermann, "Object–oriented analysis–synthesis coding of moving images", Signal Processing: *Image Communication*, Vol. 1, No. 2, pp. 117–138, Nov. 1989.

[2] J. Ostermann, "Object–based analysis–synthesis Coding based on the source model of moving rigid 3D objects", accepted for publication in Signal Processing: *Image Communication*.

[3] J. Ostermann, "An analysis–synthesis coder based on moving flexible 3D–objects", *PCS 93,* Lausanne, Suisse, No. 2.8, 17–19 March 1993.

[4] J. Ostermann, "Object–based Analysis–Synthesis Coding (OBASC) based on the Source Model of Moving Flexible 3D Objects", accepted for publication in IEEE Transactions on Image Processing.

[5] R. Forchheimer, O. Fahlander, T. Kronander, "A semantic approach to the transmission of face images," *Picture Coding Symposium (PCS'84)*, Cesson–Sevigne, France, No. 10.5, July 1984.

[6] P. Roivainen, R. Forchheimer, "Model–based coding: estimation of face motion", *Picture Coding Symposium (PCS '90)*, Boston, USA, No. 7.7, March 1990.

[7] R. Koch, "Adaptation of a 3D facial mask to human faces in videophone sequences using model based image analysis", *Picture Coding Symposium (PCS '91)*, Tokyo, Japan, pp. 285–288, September 1991.

[8] R. Rydfalk, "CANDIDE, A parameterised face", Internal Report Lith–ISY–I–0866, Linköping University, Linköping, Sweden, 1987.

[9] C.S. Choi, T. Takebe, H. Harashima, "Three–dimensional (3–D) facial model–based description and synthesis of facial expressions", IEICE Transactions Electronics and Communications in Japan, Part 3, Vol. 74, No. 7, pp. 12–23, 1991.

[10] M. Wollborn, "Object–oriented analysis–synthesis coding using prototype prediction for colour update", *Picture Coding Symposium (PCS '93)*, Lausanne, Suisse, No. 17.6, March 1993.

[11] J.A. Robinson, "Three–channel moving picture coding", COST 211^{ter} European Workshop on New Techniques for Coding of Video Signals at Very Low Bitrates, No. 4.1, December 1993.

[12] J. Ostermann, H. Li, "Detection and Coding of Model Failures in an Analysis–Synthesis Coder Based on Moving 3D–Objects," *3rd International Workshop on 64 kbit/s Coding of Moving Video*, Rotterdam, The Netherlands, No. 4.4, September 1990.