Automatic Face Location to Enhance Videophone Picture Quality

T.I.P. Trew, R.D. Gallery, and D. Thanassas
Philips Research Laboratories
Redhill, UK

E. Badiqué
Philips Kommunikations Industrie AG
Nürnberg, FRG

Abstract

New video communication and multi-media products open up a range of machine vision applications, in which the potential size of the market can justify a substantial investment in the development of sophisticated algorithms. Face location can be used to enhance the subjective performance of videophones, while still conforming with international video compression standards. This paper gives an overview of the location techniques employed, describes a real-time implementation, and presents the results of the subjective tests which confirmed the improvement in picture quality.

1 Introduction

Although many companies have been active in machine vision research for over 20 years, there are few commercial products based on this technology, other than the ubiquitous bar-code reader. This is changing with the introduction of video communication products, such as videophones, and this paper describes how techniques which solve the real-world machine vision task of face location may be applied to commercial advantage.

For videophones to become accepted as a universal business tool, as the fax machine is today, they must conform rigorously to international standards to ensure compatibility between different manufacturers' equipment. This standardisation has been achieved for videophones operating over the Integrated Services Digital Network (ISDN). There are many CCITT[1] standards which apply to the various video, speech and network access sub-systems of the videophone, but we need only consider H.261, which describes the techniques to be used for video compression.

Since all manufactures must conform to the standard, novel approaches are required to introduce product differentiation, and it is in this area that machine vision techniques are applicable. Although H.261 defines the core of the compression algorithm, so that the transmitted bit-stream can be interpreted by any decoder, additional pre- and post-processing stages are permitted. Furthermore, the standard allows high-level knowledge to be applied to control the parameters of the algorithm, so as to increase

[1] Comité Consultatif International Télégraphique et Téléphonique

BMVC 1992 doi:10.5244/C.6.51

the fidelity of parts of the picture of interest to the user, at the expense of the remainder. If applied correctly, this will increase the overall subjective image quality.

In the majority of cases, it is the user's face and, in particular, the eyes and mouth, which is of greatest interest. People are particularly sensitive to degradations in these areas and a small improvement in the objective picture fidelity, if it makes it possible to read expressions, can yield a substantial increase in the attractiveness of the product.

This paper first gives an overview of the H.261 video compression standard, since this is necessary to explain how machine vision techniques can be useful. It then describes both a simple face location system, which has been realised and interfaced to an H.261-compliant videophone, and a more sophisticated approach, which can locate individual facial features, but which has not yet been fully realised in real-time. Finally the results of subjective tests in which non-experts assessed the overall picture quality are presented.

2 H.261, the International Videophone Standard

H.261 is an international standard, developed by the CCITT Study Group XV[1] for videophone transmissions over digital networks at low bit rates (multiples of 64kbit/s). Consider that a compression ratio of 300:1 is required for the lowest rate (64kbit/s). Using current coding techniques it is not possible to achieve such a huge reduction without introducing a visible deterioration in the fidelity of the decoded image.

The basis of the H.261 coding algorithm is a hybrid of several well known techniques, and it might be described as a hybrid motion-compensated DPCM/DCT coder, where DPCM is differential pulse coded modulation, and DCT is the discrete cosine transform. Figure 1 shows a block diagram for such a system, in which the "facial area" input should be ignored at this stage. The algorithm, after initialisation, proceeds as follows. The frame store contains the image which was displayed during the previous frame period and the motion estimator, which uses block matching with 16×16 pixel blocks, termed "macroblocks", finds the best match for each block in the current frame. The motion vectors are used to displace the image in the frame store,

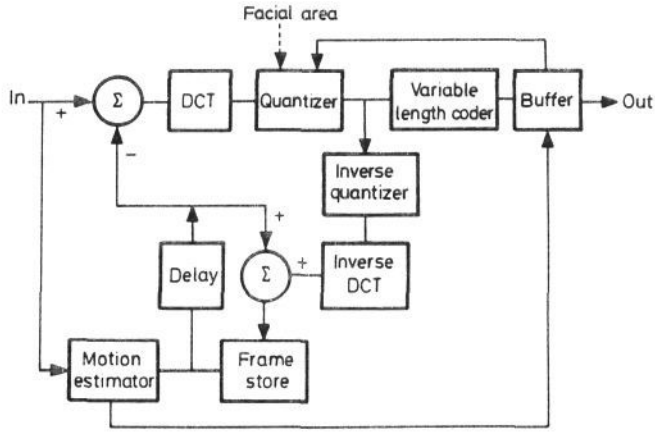

Figure 1. : Block diagram of an H.261 coder, showing how the facial area is used to modify its operation.

which is replicated in the decoder, to form the DPCM prediction. The difference between this prediction of the current image and the actual image is calculated by subtracting the two images, to give a motion compensated frame difference. This has exploited the temporal correlation within the image sequence to reduce the amount of data to be transmitted.

The next stage of the algorithm seeks to exploit the intraframe, or spatial, correlation, within the motion compensated frame difference by taking its discrete cosine transform, on an 8×8 pixel block basis. The coefficients of the DCT are quantised (introducing error), and also thresholded to discard the smaller coefficients in any block. The output of this stage is then Huffman coded [2], and fed into a buffer which matches the instantaneous data rate of the encoder to the fixed rate of the transmission channel. At this stage the data need only have error protection coding added before being transmitted. The amount of data within the buffer is monitored, and a signal is fed back to control the step size and threshold of the quantiser, which will determine the resolution and number of the transmitted DCT coefficients.

The subjective quality of the images produced by the above algorithm is dependent upon both the complexity of the image (and how suited this complexity is to the basis functions of the DCT), and also to the extent and type of motion in the image (i.e. block matching can handle 2-D planar motion, but motion involving rotation, or motion parallel to the camera axis, will reduce the correlation of the matching process), resulting in a degradation of the subjective image quality. People using videophones cannot have their movement unduly constrained, and there might be movement in the background of a typical office environment, so the problem of the degradation of picture fidelity due to motion over a significant portion of the image needs to be addressed.

In typical videophone conversations the participants are talking to each other and looking at each others faces, and are not greatly interested in the appearance of the background. Therefore, instead of using a constant quantisation step size for the whole image, the quantisation step used in the facial area can be decreased, so that more bits will be used in this area. The background will of course now receive fewer bits, and hence degrade, but, as it is not the centre of attention, the overall subjective picture quality should improve. The "face area" input, shown dashed in figure 1, leads to a modified H.261 coder, in which a mask indicating the facial region is used to control the quantiser. Figure 2 illustrates the improvement in the decoded picture quality obtained from the modified encoder, in which the step size in the facial area has been halved, compared with the standard encoder.

Having established how the videophone coder will be modified, it is now necessary to devise a machine vision system which is capable of locating the user's face automatically in a normal office environment.

3 Low-level Face Location

Initial experiments concentrated on enhancing scenes in which a single person was present (head-and-shoulder scenes). One can suppose that an especially high quality will be expected when two speaking partners have an eye-to-eye conversation. In these situations, it will be important to be able to have a clear image of the partner's face in order to help in interpreting his expression. The quality of the facial image will be far less critical in situations where several people are in the field of view and the scene activity is high. The first step, prior to the realisation of a facial area recognition

(a) (b)

Figure 2. : A comparison of the decoded pictures from standard (a) and modified encoders (b).

algorithm which is simple enough for integration within a codec, is the automatic identification of scenes of one person's head and shoulders.

3.1 Recognition of Head-and-shoulder Scenes

The recognition algorithm is based on the detection of the motion within the scene. The different steps are detailed below:

3.1.1 Generation of an Object Silhouette Based on Image Differences

To generate an object silhouette characterising the moving object, the image difference is built over blocks of a given size (8×8 or 16×16). The decision as to whether a block belongs to the moving object or not is then taken based on the temporal activity within the block, characterised by the conjunction of two criteria. The first criterion is the sum of the absolute value of the image difference. The second criterion is the number of points within a block, for which the absolute value of the image difference is larger than a given threshold.

3.1.2 Spatial and Temporal Filtering of the Object Silhouette

Due to a number of factors, the raw silhouette generally cannot be used for analysis without being filtered. Spatial and temporal filters are used in order to improve the smoothness of the silhouette.

As a first step, the silhouette is passed through a non-linear spatial filter. The eight neighbouring blocks of a given moving block are analysed and, according to their status, the moving block is set to "fixed". As a result, most of the isolated "moved" blocks appearing in the background due to camera noise are suppressed and the analysis of the silhouette is simplified.

The spatial filter is followed by temporally smoothing the motion silhouette. In this step the temporal stability of the bitmap is improved by ORing the status bit for a given spatial position over several frames.

3.1.3 Transformation of the Silhouette into a Feature Vector

After filtering the motion silhouette, a 1-D feature vector can be extracted and its topological characteristics can be analysed. The most important feature is the width of the silhouette in each row of blocks. This feature in a typical head-and-shoulder scene will usually have two characteristic regions, corresponding to the head and the shoulders. In order to ensure a recognition performance as size-insensitive as possible, the feature vector is normalised between 0 and 1. Together with the feature vector, an average value of the position of the vertical axis, and also of the top of the moving object, are calculated. These are stored to be used later during the localisation of the head in the case of a head-and-shoulder scene.

3.1.4 Training of the One-dimensional Pattern Associator

Through the transformation described above, the problem was reduced to the analysis of the 1-D vectors. This type of problem can be solved with the help of a pattern associator[4], trained on features contained in a training set.

Let us assume that N typical head-and-shoulder silhouettes are recorded during training and their 1-D feature vectors are obtained. The vectors are described as f_i; $i \in \{1, \ldots, N\}$. In order to be able to differentiate between the "true" head-and-shoulder and false silhouettes we also need feature vectors representing "false" (i.e. non-head-and-shoulder) silhouettes. We generate these "false" silhouettes by, for example, inverting the silhouette about its horizontal axis. We add these "false" 1-D feature vectors to the training set and call them f_i; $i \in \{N+1, \ldots, 2N\}$. The constraint equation determining a filter H is then:

$$\mathbf{f}_i^T.\mathbf{H} = C_1 = 1; \quad i \in \{1, \ldots, N\} \tag{1}$$

$$\mathbf{f}_i^T.\mathbf{H} = C_2 = -1; \quad i \in \{N+1, \ldots, 2N\} \tag{2}$$

The filter **H** is determined iteratively with the help of a delta-rule training algorithm. It is written as follows (at the t^{th} iteration):

$$\begin{aligned} \mathbf{H}_t = (\mathbf{f}_i^T\mathbf{H}_t - C_k)\mathbf{f}_i \quad & k=1; \; i \in \{1, \ldots, N\} \\ & k=2; \; i \in \{N+1, \ldots, 2N\} \end{aligned} \tag{3}$$

For successful training it is necessary that a training set contains 1-D feature vectors which are representative of typical videotelephony situations, with different subjects and various subject-camera distances, etc.. It is also possible to generate a training set from whole sequences, using either simple solutions, such as averaging the different silhouettes, or more sophisticated ones, such as principal component analysis.

3.1.5 Filtering the Feature Vector

Having trained the pattern associator off-line, it is then combined with the current feature vector by a simple dot-product operation, as in equation (1). A head is detected if the resulting value exceeds a threshold.

A number of temporal consistency rules can then be used to smooth the decision and ensure stable head-and-shoulders recognition.

3.2 Localisation and Tracking of the Facial Area

Having established that a head-and-shoulders silhouette is present in the scene, it is necessary to localise the area which is to be enhanced. Since the H.261 standard only allows the quantisation step size to be changed on the boundaries of 16×16 pixel macroblocks, the face is represented as a rectangle containing a whole number of macroblocks. Only limited accuracy is required because of this coarse scale and experiments have shown that only three different rectangle sizes are required to cover the range of head sizes observed in normal videotelephony situations. The rectangle size is a function of the total silhouette size and it is positioned using the vertical axis and top-of-object position determined during the silhouette analysis.

For those frames in which an acceptable face cannot be found, possibly because of complex background activity, the facial area can be tracked from previous frames. A tracking system has therefore been developed which segments the motion vector field to identify areas which are moving coherently. Those segments which correspond to the facial area are identified and the centroid and mean motion vector for this group are calculated. This vector is used to project the centroid into the following frame. The segmentation of the motion vectors is repeated for that frame and a group of segments is grown around the projected centroid, observing spatial and temporal coherence criteria, according to the hypothesis that the motion and position of the head change smoothly in consecutive frames. If these criteria cannot be met then the system takes into account past information about the position and location of the head, as it was last detected in the previous frames. This assumes that the head has not moved enough to generate a coherent vector field, hence its position and motion characteristics can be recovered using past motion history.

This system can be used as a general purpose object tracker provided that the initialisation stage is changed accordingly. Initial simulations, with no code optimisation, run on a 12MIP RISC processor at the rate of 6 frames/sec.

4 Overview of Feature-Based Face Location

Although the face location system described in section 3 operates successfully for most of the time, and has been realised in real-time, the silhouette is very noisy, so that the system fails occasionally, and is coarse, so that the size of the face cannot be determined precisely. Also the positions of facial features, required for other enhancements to the videophone, cannot be obtained.

An alternative, bottom-up approach has been developed in parallel with the previous system. This attempts to identify features in the scene which may be the eyes or the mouth, and then finds triplets of these with the correct geometry to form a complete face. The principles of this approach have already been described[3], but there have been several extensions since that time which have made the technique more robust.

The principle of the approach is that potential faces are located from the current image at regular intervals, determined by the processing power available, and the probability of each of these potential faces being a true face is estimated. The potential faces are tracked through subsequent frames, allowing the probabilities from the individual frames to be smoothed with a non-linear temporal filter. The potential face with the highest smoothed probability is selected. The following sections describe this procedure in more detail.

4.1 Initial Face Location

The technique initially analyses the images to find spatio-temporal features which might be an eye or mouth. The criteria for selecting these points are not very stringent since the scale of the face is unknown initially. Triplets of these points are then considered, using rather strict geometric constraints, to determine which of these groups might form a face. These constraints are set so that a face will only be accepted when the line of sight of the user is close to the camera.

It is now necessary to assign a probability to each of these triplets. This is achieved by assessing the image around the vertices which we postulate may be the eyes. For each triplet, this area is normalised, both geometrically using the size and orientation of the triplet, and for illumination, and is then applied to a perceptron. The perceptron has been previously trained on true eyes and on other features which the low-level analysis has incorrectly identified as potential eyes. Note that it is not possible to recover the full 3-D pose of the face from the triplet geometry obtained from a single camera. The training regime accounts for this ambiguity by repeating each true face several times, to have an equal number of true and false training examples, but adding a random value to each of the geometric normalisation parameters. The probability is assigned to the potential face by using the probability distribution function of the perceptron for the eyes of example faces.

4.2 Tracking and Temporal Filtering

Having instantiated potential faces, as described above, they are tracked through subsequent frames. Tracking allows triplets instantiated in different frames to be associated, so that their probabilities can be smoothed, reducing the number of occasions on which spurious triplets are selected. It also allows triplets to persist, even if feature points are not found at all of the vertices in every frame. These propagated triplets are subjected to the same types of geometric tests described above, but now the criteria are relaxed so that the face will not be discarded, even when the face is turned away from the camera.

Although it would be convenient to use the motion vectors available from the H.261 codec, these vectors are not ideal for this process since they are computed to project the current frame back into the previous frame to satisfy the requirements of the inter-frame coding mode of the H.261 compression standard. In contrast, triplet tracking requires vectors which are estimated looking forwards from the current frame. If the sense of the vectors from the codec are reversed then they can give rise to ambiguities caused by covering and uncovering, and more reliable results can be obtained by using a separate tracking system. Satisfactory results can be obtained using the technique proposed by Seeling [5,6], which also explicitly accounts for lighting changes and occlusion.

4.3 Face Location

For simple sequences, with only a single face in view, the face can be located by selecting an area around the triplet with the highest smoothed probability. For more complex sequences with several people, a higher level mechanism is required to select the individual who should be located. Currently, simple heuristics, such as selecting the person closest to the centre of the screen, are used, but more sophisticated approaches might select the person who is currently speaking.

5 Experimental Hardware and Subjective Tests

The low-level analysis algorithm described in section 3 was implemented in real-time and was extensively tested as a standalone system, as well as being connected to an H.261 coder. The purpose of the real-time experiment was to complement earlier computer simulations and to find out in "real life" situations whether the recognition and improved coding of the facial area in an H.261 codec did lead to an improvement in the overall subjective image quality.

5.1 Hardware Description

The facial area recognition system consists of a videophone module (camera and display), an A/D converter, a block-based motion analysis board and a PC for the analysis and display of the facial region, as shown in figure 3.

The data from the motion analysis board (a 36×44 bitmap corresponding to 8×8 blocks) is sent to the PC through a parallel port. The data is displayed on the PC screen using one pixel per block to guarantee a real-time display. The data is filtered and analysed with a 1-layer neural network and, if a head-and-shoulders scene is recognised, a rectangular box of variable size corresponding to the facial region is superimposed onto the displayed data. The vertices of this rectangle are sent to the codec through an RS232 serial port, causing the codec to code the blocks within that region with the quantisation step size set to a fraction of the value determined by the buffer regulation. In order to avoid buffer overflow, the rest of the scene is coded with a step size which has been correspondingly increased.

In order to facilitate experimentation and optimisation, both the parameters of the analysis program on the PC and the coefficient multiplying the step size in the codec can be changed interactively. The coefficient can be varied between 0.1 and 1.0 in 0.1 steps by hitting a key on a terminal keyboard. The optimal range for this coefficient was found experimentally to be between 0.3 and 0.5.

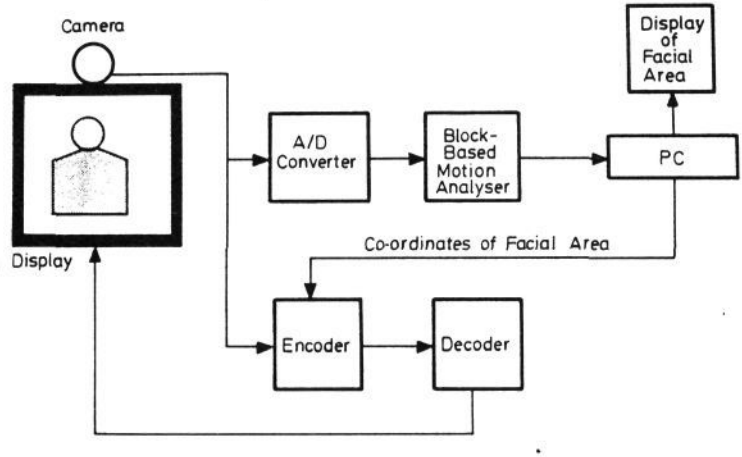

Figure 3. : Block diagram of the sub-systems in a videophone incorporating face location.

5.2 Experimental Results and Subjective Tests

The first results of the real-time test confirmed the indications given by computer simulations. The improved coding of the facial region produced videophone images which tend to have much less blocking in areas of high subjective significance, such as the eyes and mouth. The blocking in such areas is often strongly perceived as being annoying because it inhibits eye contact or lip reading. With the improved coding of the facial area, the image of the speaking partner is reproduced in a way which is more acceptable to the untrained observer.

In order to objectively evaluate the significance of the improvement, more than 30 untrained and unaware observers were asked to judge the difference in image quality between standard coding and adaptive coding. The observers were either asked to sit in front of the camera and to observe their own coded image (self-image mode) or to observe the decoded image of a second person on a monitor (conversation mode). Some of the observers were given a telephone receiver and were asked to simulate a telephone conversation. They could toggle between the two modes, but they were not aware what differences were expected in the two modes, nor which of the modes was currently selected. The observers were then asked whether they could detect any difference between the modes and, if so, to rank the picture quality between them. The experiment was recorded on videotape for later analysis.

The experiment lead to the following observations. One half of the group of observers (18 people) were only asked to observe the scene in self-image mode. The majority of the observers (14 persons) characterised the improvement as "small" to "significant". The other four persons either saw no difference (generally they had not moved much during the test) or a very small improvement without significance. The other half of the group (19 persons) were asked to first observe a self-image and then an image in conversation mode. In this case all the observers made the same pattern of judgements. When asked to evaluate a self-image, they all saw a small to significant improvement. They all described the improvement in the conversation mode images as "significant" to "very significant". This showed that the quality improvement in the facial area is perceived more strongly when the observer is communicating with another person than when he is observing his own self-image. Self-image observers tend to concentrate less on the face and to pay more attention to other parts of the body or to the background. Conversation-image observers tend to concentrate on the eyes and mouth of the speaking partner, thus being more sensitive to quality improvements in these regions. As expected, the quality improvements are most appreciated when they are most needed, that is during a telephone conversation.

6 Conclusions

New video communication and multi-media products open up a range of machine vision applications, in which the potential size of the market makes it worthwhile to make a substantial investment in the development of sophisticated algorithms to obtain an advantage in the market. The US videoconferencing market is expected to be worth $8.3 billion by 1995[7] and any developments which can increase a company's market share are clearly worthwhile. At the same time, the advent of very fast video processing devices, which can be programmed in high level languages, such as the Philips LIFE VLIW device[8], make it possible to realise complex algorithms in real time and at low cost, with short development times.

The subjective tests in this paper demonstrate the potential for machine vision to enhance the performance of an H.261-compatible videophone, and research is continuing to use the capability to locate the user's head to enhance other aspects of the equipment.

References

[1] CCITT Recommendation H.261, "Codec for Audiovisual Services at n x 384kbit/s", *CCITT IXth Plenary Assembly*, 1989.

[2] D.A. Huffman, "A Method for the Construction of Minimum Redundancy Codes", *Proc. IRE*, **40** (10), pp.1098-1101, 1952.

[3] E. Badiqué, "Knowledge-Based Facial Area Recognition and Improved Coding in a CCITT-Compatible Low-Bitrate Video-Codec", *Picture Coding Symposium*, Boston, March, 1990.

[4] J.L. McClelland and D.E. Rumelhart, "Explorations in parallel distributed processing", MIT Press, 1986.

[5] G.C. Seeling, "Tracking 3-D Moving Objects", Imperial College Msc Thesis, Philips Research Laboratories internal report, Redhill, UK, September, 1990.

[6] "Tracking a Moving Object" European Patent EPA 0474307, March, 1992.

[7] M. Nakamoto, "Phones with a View to Profit from Vision", *Financial Times*, London, 28th April, 1992.

[8] G.A. Slavenburg, A.S. Huang and Y.C. Lee, "The LIFE Family of High Performance Single Chip VLIW's", *Hotchips III*, Palo Alto, California, 1991.

Acknowledgements

Sincere thanks are due to J. Hempel and B. Krupa (PKI) for their great help in realising the real-time recognition system. The subjective tests would not have been possible without the help of Mr. Kummerfeld and Mr. Wolf (Daimler Benz Research) who kindly reprogrammed a DB codec for this purpose. Part of this work was realised with the financial support of the European Community RACE-HIVITS project.