Multi-Modal System for Locating Heads and Faces

Hans Peter Graf[1], Eric Cosatto[1], Dave Gibbon[1], Michael Kocheisen[1], Eric Petajan[2]
[1]AT&T Labs, Room 4G 320, Holmdel, NJ 07733, USA; hpg@research.att.com
[2]Lucent Technology, Bell Laboratories, Murray Hill, NJ

Abstract

We designed a modular system using a combination of shape analysis, color segmentation and motion information for locating reliably heads and faces of different sizes and orientations in complex images. The first of the system's three channels does a shape analysis on gray-level images to determine the location of individual facial features as well as the outlines of heads. In the second channel the color space is analyzed with a clustering algorithm to find areas of skin colors. The color space is first calibrated, using the results from the other channels. In the third channel motion information is extracted from frame differences. Head outlines are determined by analyzing the shapes of areas with large motion vectors. All three channels produce lists of shapes, each marking an area of the image where a facial feature or a part of the outline of a head may be present. Combinations of such shapes are evaluated with n-gram searches to produce a list of likely head positions and the locations of facial features. We tested the system for tracking faces of people sitting in front of terminals and video phones and used it to track people entering through a doorway.

1. Introduction

Many algorithms for identifying faces in images have been described in the literature [see e.g. 1]. Such algorithms tend to work well over a limited range of conditions, but often fail when exposed to a 'real world' environment, where lighting conditions, camera characteristics or other scene parameters vary. For a head location system to be perceived as truly non-intrusive by the observed people, a free motion of the heads has to be permitted. This results in large variations of the heads' sizes and orientations. To handle such a large range of conditions efficiently, we combine the information of three channels: shape, color and motion. This has resulted in a robust face and head location system suited for such applications as tracking people for surveillance purposes, model-based image compression for video telephony [2], and intelligent computer-user interfaces. It was tested in several 'natural' environments, for example, it was set up in the hallway of a building for capturing the heads of people entering through a doorway. In other tests a camera, installed on a workstation, was tracking heads as well as individual facial features of people sitting in front of the terminal. In both these situations no special lighting was installed, and the people who were observed could move freely.

Having multiple channels producing results, the main challenge is to combine all this information efficiently. Various attempts have been described in the literature to improve accuracy and robustness of recognition systems by combining the results of multiple algorithms and classifiers. Often, several different classifiers evaluate an object independently and then the results are combined in a final step, for example by voting [3]. Other techniques combine the results of different classifiers with weights that take into account the error rates of each of the classifiers. By integrating the final evaluation of the different classifiers into the training procedure, the accuracy of a system can often be increased further. The approach taken here is to combine the results of the different channels as early as possible. All three channels produce the same intermediate representation where areas, that may contain facial features or head outlines are marked. A single classifier, is then evaluating combinations of such areas.

We want to combine the different channels of information not only to improve the robustness and accuracy, but also to accelerate the speed. Often simple and fast algorithms, such as color or motion analysis provide enough information to track faces or facial features. Hence, a lot of time can be saved if only the appropriate channel is chosen. But in order for such algorithms to work reliably, the system needs to have a high confidence that one channel can provide adequate information. This is

0-8186-7713-9/96 $5.00

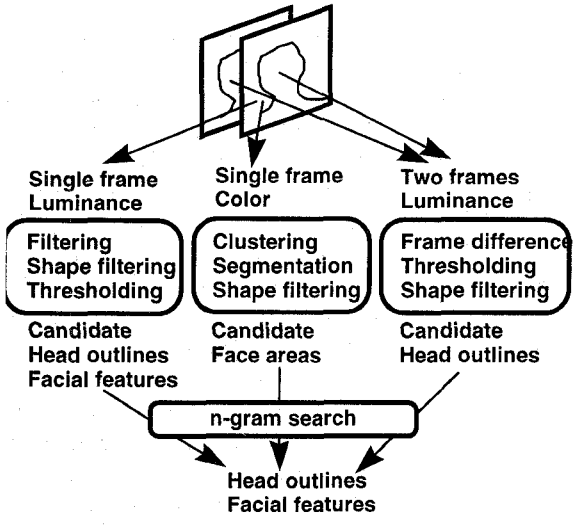

Figure 1: ***Overview of the sequence of algorithms used for finding faces and facial features.***

achieved by first doing a thorough analysis of the whole image where all three channels are evaluated, and then the decision is made which channel is suited best for further tracking.

For example, when a face is tracked, often a simple color segmentation provides accurate results if there is a good contrast between skin colors and those of the background. If, on the other hand, the background colors are similar to skin colors, finding the accurate location of the face is more difficult, and color segmentation is likely to fail. Hence, key to an accurate analysis is, to identify the best strategy, and in order to save computation, it has to be found as early in the analysis as possible. Starting with a full evaluation of all three channels provides the robustness and allows to compare the results of all of them. For each channel a measure of confidence is computed and if that is high enough for the color channel, then we track a face using only color segmentation for several frames. During this time the search for facial features is limited to the area identified as face by the color segmentation. After about 10 frames the result of the color segmentation is confirmed by invoking the motion analysis. In this way, during the majority of the time only the very fast color segmentation process has to be calculated plus a reduced version of the shape analysis, saving a lot of computation as compared with the full analysis of the three channels.

Figure 1 shows an overview of the data flow and the type of algorithms applied. Section 2 describes the shape representation of the images, while section 3 is focusing on the color analysis, section 4 describes the motion analysis, and section 5 briefly discusses how the results of the different channels are combined.

2. Shape Analysis

The shape analysis tries to find outlines of heads or combinations of facial features that indicate the presence of a face. It uses luminance only, and therefore can even work with cheap monochrome cameras. For frontal views of faces, we first identify candidate areas for facial features, and then we search for combinations of such areas to find the whole faces. In images with a low resolution, individual facial features may not be distinguishable, or a person may turn away from the camera, so that only the back of the head is visible. In such cases we depend on finding the outline of the head.

Several head location schemes based on shape analysis have been proposed by other researchers. Local detectors have been used to identify areas of facial features [4] or the outline of the head [5].

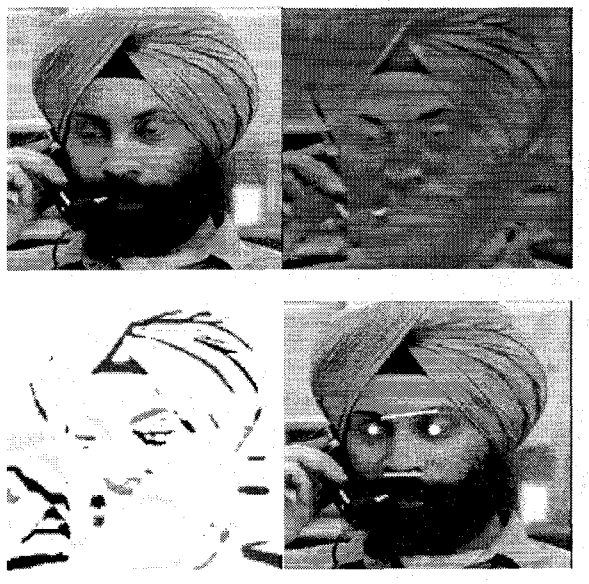

Figure 2: ***Example of the facial feature detection process. The top-left image is the original, the top-right image has been filtered to select a range of spatial frequencies and sizes. The bottom-left shows the image after adaptive thresholding where areas of interest have been marked with connected areas of gray pixels. The bottom right image shows the best combination of facial features that could be identified: eyebrows, eyes and nostrils.***

Other techniques use trained filters for detecting whole faces and heads [e.g. 6].

The key element of our shape analysis is an intermediate representation of the images from which facial parts or head outlines can be located easily. The feature detection scheme described here is a generalization of the process presented last year at this workshop [7]. We expanded the analysis by taking more different facial features into account, adding a separate search process for head outlines and make the whole process fully trainable.

An image is first transformed by two filters, the first one is a band-pass filter selecting a range of spatial frequencies. The second one is tuned to detect a range of sizes of a simple shape such as a rectangle or an ellipse. These processing steps reduce variations due to different lighting conditions and enhance areas of facial features or head boundaries.

Facial features exhibit intensity variations and hence their appearance can be emphasized by selecting a band of spatial frequencies. Once filtered for spatial frequencies, the image is filtered for certain shapes. This is accomplished by convolving the image with a rectangle or an ellipse. In this way areas of high intensity that are larger than the structuring kernel are emphasized while smaller areas are reduced in intensity.

Figure 3: *Examples of head locations, as well as eye and mouth locations, found with the shape analysis. In the face on the left only the search for the head outline succeeded since the search for facial features is limited to a head tilt of +-30 degrees.*

After the filtering operations, the image is thresholded with an adaptive thresholding technique. The goal is to identify the positions of individual facial features with a simple connected component analysis. If the threshold is chosen properly the areas of the prominent facial features, such as eyes, mouth, eye brows, and the lower end of the nose are marked with blobs of connected pixels that are well separated from the rest. Examples of this representation are shown in Figures 2 and 4. One can then locate the positions of a face by looking for appropriate combinations of these blobs. For finding the outline of the head, the images are treated in the same way, but now vertically as well as horizontally extended regions of high spatial frequencies are filtered out.

Once candidate facial features are marked with connected components, combinations of these features, that could represent a face, have to be found. This is done with an 'n-gram' search. First the shape of each individual connected component is analyzed and those that can definitely not represent a facial feature are discarded. Then combinations of two connected components are tested whether they can represent a combination of two facial features, for example an eye pair, eye brows, or an eye and a mouth. In the next step triple combinations are evaluated, etc. In each of these steps the connected components are evaluated with small classifiers that take as their inputs the sizes, ratios of distances and orientations of the connected components. The facial features being considered are the eyes, eye brows, nostrils, mouth, and chin groves.

The search for the head outline proceeds in the same way. The first scan through the connected components selects those that could represent right or left vertical boundaries of a head. Then combinations of right and left edges are examined and finally combinations of vertical and horizontal edges. The head outline is approximated with an ellipse, and the coverage of an ellipse by connected components is taken as a measure of the goodness of the fit. If results from the other two channels are available, they are included in the n-gram search. This is explained in more detail in section 5.

The computation of the n-gram search grows exponentially with n, the number of different components taken into account, and hence it is potentially costly to compute. However, by using this hierarchical search and eliminating components with low scores in each step of the search, this computation can be kept very fast. In fact, the computation for the whole shape analysis is dominated by the time for the band-pass filtering and the shape filtering. On a PentiumPro (150MHz) the whole shape analysis of an image with a size of 360x240 pixels takes less than 0.5 sec.

Several parameters are required for this process, namely: The cut-off frequencies of the band pass filter, the size of the structuring kernels for the shape filtering and the thresholds for binarizing the

results. The optimal values for all these parameters are determined with a fully automatic training procedure, using 100 images of 25 different people. In the training set the positions of the eyes, the left and right end points of the mouth, and the lower end of the nose are measured by hand.

Then the images are treated as described and the sizes of the connected components in the area of the facial features are measured. For an automatic optimization of the parameters a quality measure of the following form is used:

$$s = 100 - a*(x - x0) - b*(w - w0)$$

S: Quality of the marking of the feature
x: position of the connected component
$x0$: The desired position
w: width of the connected component
$w0$: desired width
a, b: scaling factors

For training we do an optimization of each parameter independently by scanning one parameter over its whole range of values while the other parameters are kept constant. Figure 4 shows an example of such a sweep, where the threshold for binarizing the filtered image is tuned.

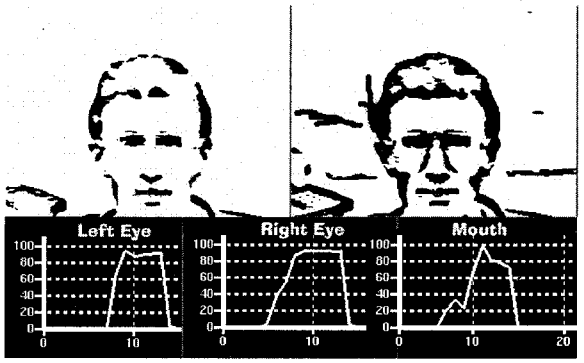

Figure 4: *Marking areas of interest for identifying the positions of facial fatures. The top two images have been filtered for spatial frequences and sizes and are binarized with two different thresholds. The bottom diagrams show the quality index for marking left eye, right eye and mouth, respectively. The vertical axes show the quality index, the horizontal axes the threshold value. The top left image is binarized with threshold 12, close to optimal, and each of the three features is clearly marked with a separate connected component. The top right image is binarized with a threshold that is too low (coresponds to 7 in the diagrams). In this case the eyes are not marked individually and will not be selected for further analysis.*

When the parameters are trained properly the facial features are spotted correctly over a wide range of scales and conditions. For example, eye regions are found, regardless whether the eyes are open or closed. The same is true for mouths. Whether the mouth is open or closed, whether the teeth are visible or not, has little influence on the ability of the described technique to mark the correct area. A big advantage of this approach is that a wide range of sizes can be covered with one set of parameters. Filters designed for detecting whole faces or heads tend to be very scale sensitive and many search scans need to be performed to cover a reasonable range of sizes. Our technique handles typically a range of more than a factor of two in head size with a single set of parameters, i.e. the height or width of heads can vary by more than a factor of two. On a test set of 120 images of 30 different people in poses typical for someone sitting in front of a terminal, head positions are found correctly in 90% of the cases. Eyes and mouth are located correctly in 88% of all images.

3 Color analysis

Several studies have been published, where color alone was the feature for identifying the area of a face. In our experience, color information is an efficient tool for identifying facial areas and specific facial features if the system can be calibrated properly for particular conditions. However, these calibrations can usually not be transferred to different cameras and to strongly varying conditions in the illumination. Studies with several thousand photos showed [8] that skin colors can vary a lot and often are indistinguishable from similar background colors. Therefore, we use color only in combination with shape and motion analysis, where we can calibrate the color space first.

For finding a whole face, the color space is clustered with a leader clustering algorithm, where one or two cluster centers are initialized to skin colors of a part of the face identified by the shape analysis. As color space we chose normalized rgb values [r = R/(R+G+B), g = G/(R+G+B), b = B/(R+G+B)] in order to minimize the algorithm's dependence on luminance. Dark pixels (R+G+B < 30) were set to zero to avoid instabilities caused by the normalization. A few of the processing steps are shown in Figure 5. After skin colors have been identified with the calibration and the clustering process the image is thresholded in order to locate the area of the face.

When we are looking for whole faces, color information is only used to identify larger areas, and hence the picture is subsampled to 40 x 30 pixels using bilinear interpolation. After binarization each segment in the image is analyzed for its shape and size to determine whether it can represent a face or not. As is indicated by Figure 5, faces are often the dominating connected components in the image and the face position can easily be identified. The time required for the color analysis, once the calibration is done, is around 10 milli-seconds on a 90 MHz Pentium.

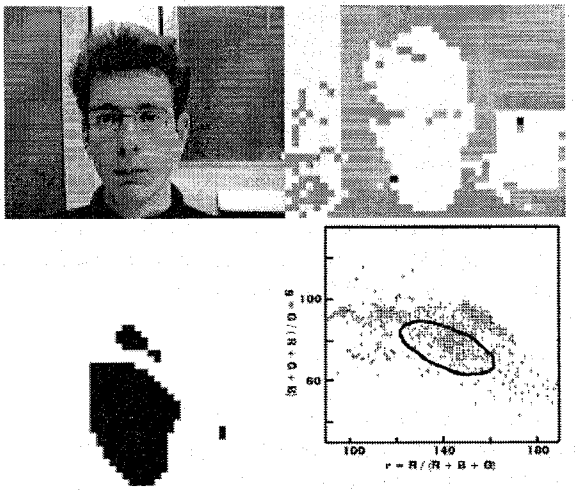

Figure 5: *Example of the color analysis. The top left image is the original and the top right the downsampled representation, where different hue values have been transformed into different gray levels. The bottom left image is the segmented image, where the colors inside the ellipse of the diagram at the bottom right were taken as skin colors. The bottom right diagram shows the distribution of the image colors in the r-g plane. Many of the background colors are very similar to the face colors, and without calibration the face could not be distinguished from the background.*

4. Motion analysis

If multiple images of a video sequence are available, motion is often a parameter that is easily extracted, offering a quick way for locating objects, such as heads. The first step in the algorithm is to compute the absolute value of the differences in a neighborhood (typically 8x8 pixels) surrounding each pixel. When the accumulated difference is above a predetermined threshold T, we classify the pixel as belonging to a moving object. T is typically set at 1.5 times the temporal noise standard deviation, times the number of pixels in the neighborhood.

By applying the threshold to the accumulated difference rather than the individual pixel difference, we gain two benefits: First, T can be expressed with increased precision. Second, the neighborhood processing has an effect similar to morphological dilation. This helps fill small gaps that occur in areas where the moving object has similar pixel values to the background. We have found the technique to be very effective on a wide variety of cluttered background scenes

Areas of moving objects are analyzed by using a contour following algorithm to extract the region boundaries. For each region, the contour is

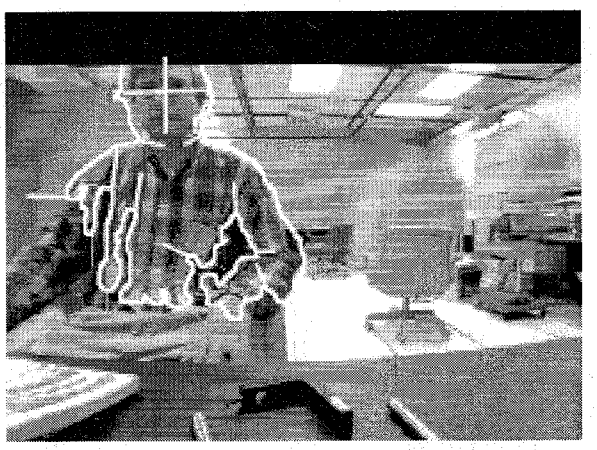

Figure 6: *Example of the motion analysis. The white outline shows the boundary of a moving object. The cross marks the center of the head.*

smoothed, and the curvature of the contour is calculated. Feature points are identified along the contour at points of local extrema of the curvature.

The set of feature points for each region are compared to a model set of features corresponding to a head and shoulders shape. If a match is found, the head center coordinates are determined by calculating the mean value of the contour data for the portion of the contour that corresponds to the head. The size of the head is estimated as the mean distance from the head center to the contour. The temporal correlation of head center and size estimate is analyzed over several frames to identify spurious matches. Since only the outline of the head is analyzed, both, front and back views and usually also side views of heads are found.

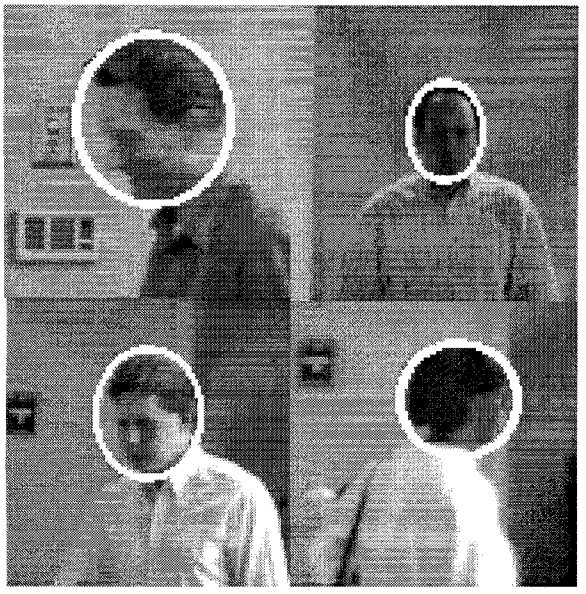

Figure 7: ***Examples of heads located with the motion detection algorithm.***

The technique analyzes a frame in less than 30 milliseconds. Installed in a hallway to detect people entering through a door, it located over 94% of all people correctly.

5. Combining the channels

A common intermediate feature representation is produced by all channels. From these feature representations connected component analysis extracts lists of shapes marking areas of interest. In this way all channels are combined seamlessly and the n-gram search described in section 2 can be applied to all of them.

The classifications are based on one or a few head models chosen to represent the expected situations. The model defines all the size parameters required for the classifications, as well as the order of the searches. To avoid a combinatorial explosion when exploring shape combinations, a greedy search is done, and it is important to establish a proper search order. The order of the searches is based on a maximum entropy measure and is determined in a training procedure.

For example, for frontal views the model is generated from a training set of 35 people looking into the camera. On this set the positions of the eyes and the eye pairs are marked correctly in over 90% of these cases. Moreover, the number of shapes that remain to be evaluated has dropped from 65 before the eye-pair search to 3.5 on average after. Hence, the eye pair search reduces drastically the number of shapes that have to be taken into account for further analysis and eye-pairs can be found reliably. Eye-pairs actually get the highest score of all two-element features. This means, that the search will start looking for eye pairs. Other features and feature combinations are classified in the same way, and an order of the searches is established.

The controller of the whole system is a state machine that tries to maintain a high confidence score for the location of the faces and facial features by invoking the appropriate combination of channels. Experiments in various environments show a great improvement in robustness of the system as compared to that of individual channels. Furthermore, by selecting only one or two channels for a part of the time the computation can be kept low. This system, in fact, represents a considerable improvement in speed and accuracy compared to a previous system with only the shape analysis channel. The accuracy for finding the head positions increased from 90% to 95% and for eyes and mouth from 88% to 93%.

References

[1] M. Bichsel (ed.), Proc. Int. Workshop on Automatic Face- and Gesture-Recognition, Zurich, 1995.

[2] T. Chen, H.P. Graf, K. Wang, Lip Synchronization Using Speech-Assisted Video Processing, IEEE Signal Processing Letters, 2/4, 1995, pp. 57 - 59.

[3] T.K. Ho, *A Theory of Multiple Classifier Systems*, Dissertation, SUNY Buffalo, 1992.

[4] M.C. Burr, T.K. Leung, P. Perona, *Face localization via Shape Statistics*, Proc. Int. Workshop on Automatic Face- and Gesture-Recognition, M. Bichsel (ed.), Zurich, 1995, pp. 154-159.

[5] A. Jacquin, A. Eleftheriadis, *Automatic location tracking of faces and facial features in video sequences*, Proc. Int. Workshop on Automatic Face- and Gesture-Recognition, M. Bichsel (ed.), Zurich, 1995, pp. 142-147.

[6] R. Vaillant, C. Monrocq, and Y. LeCun, *Original Approach for the Localization of Objects in Images*, IEE Proc. Vis. Image Signal Process. Vol. 141(4), pp. 245 - 250, 1994.

[7] H.P. Graf, T. Chen, E. Petajan, E. Cosatto, *Locating Faces and Facial Parts*, Proc. Int. Workshop on Automatic Face- and Gesture-Recognition, M. Bichsel (ed.), Zurich, 1995, pp. 41 - 46.

[8] M. Kocheisen, *Head Detection on Low-Resolution Color Photos*, Proc. Workshop Machines that Learn, Snowbird, UT, 1996.