

Machine Vision of Faces and Facial Features

Hans Peter Graf, Eric Cosatto, and Gerasimos Potamianos

AT&T Labs-Research, 100 Schulz Drive, Room 3-134, Red Bank, NJ 07701-7033, USA

{hpg,eric,makis}@research.att.com

Abstract

The recognition of faces and facial features is of great practical significance and provides a challenging proving ground for vision algorithms. Present systems tend to work well under closely controlled conditions, yet often fail miserably when exposed to variations in lighting conditions or to different head poses and different complexions.

Biological vision systems process multiple data streams in parallel and combine their results, to obtain robust recognition with a high accuracy. Increasingly, such strategies are being tried in modern machine vision systems. We then face the difficult question, at what level and how to combine the information from the different channels.

We describe here a system for face analysis that uses shape, color, and motion analysis, combining the data at an early stage of the feature representation. An application to audio-visual speech recognition demonstrates how visual and acoustic features are combined to improve the accuracy of speech recognition.

1. Introduction

Many algorithms for recognizing faces in images and for analyzing facial features have been described in the literature [see e.g. 1,2,3]. Most algorithms tend to work well over a limited range of conditions, but often fail when conditions vary, such as the lighting, the camera characteristics or the head orientations. To handle a wider range of conditions efficiently, more and more machine vision systems apply multiple algorithms and then combine the results of the different channels. One of the main challenges for such multimodal systems is, to integrate all the different results in one coherent interpretation of an image.

One possibility is to process the data in the different channels of analysis independently and combine them only at the end with one final classification. Such a late integration is attractive due to its simplicity, but misses a lot of information that can be extracted from combining features of the different channels at an earlier stage. Moreover, a look at biological vision systems suggests that an earlier integration of the information may be advantageous.

Biological vision systems use multiple types of analysis and can extract meaningful information from the combined results. The primate vision system consists of a large number of coupled modules, indicating that their information is combined, at least partially, at an early stage of the analysis. Recent studies on the combination of visual and acoustic data suggest that even these very different modalities are combined at an early stage [4,5].

In this paper we demonstrate how combining the features of three different channels of analysis lead to a robust machine vision system for locating facial features, and for measuring the precise shapes of the lips. Finding the location of a facial feature accurately requires the discrimination of fine nuances in texture and color. For lip reading the lip edges should be determined with very high precision. Often there is hardly any contrast in color or texture between the lips and the surrounding skin, which makes locating the outline of the mouth difficult. When uttering plosives, such as 'b' or 'p', the lip motion is so fast that mouth shapes in subsequent frames, recorded at 30 Hz, can differ substantially. One can therefore not rely on finding a shape similar to the one in the previous frame, which makes the analysis more challenging.

These problems have led other researchers to have the speakers wear marker points or lipstick providing a high

contrast with the surrounding skin. Some people use head-mounted cameras to maintain constant geometric arrangements. More recently lip reading experiments were done with less intrusive observation techniques. Techniques for analyzing mouth shapes are often based on edge extraction or color segmentation plus deformable templates [see e.g. 6,7].

So far, automatic lip reading has been demonstrated only for individual speakers or for a small number of different speakers. No speaker independent audio-visual speech recognition has been reported. This is due to difficulties of finding facial features precisely enough in many different people and due to the lack of suitable databases.

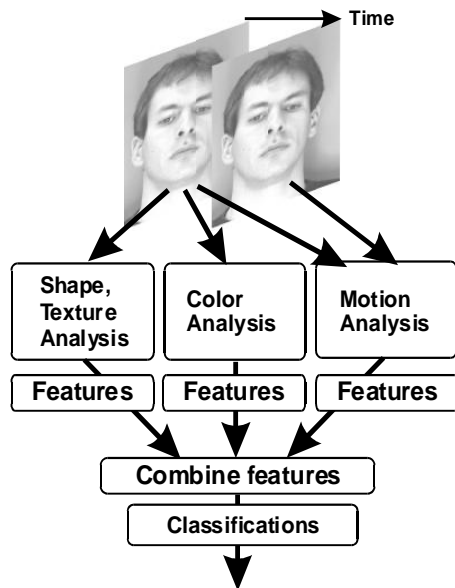


Figure 1: Schematic overview of the image analysis process.

The present investigation is aiming at speaker-independent audio-visual speech recognition. We present in the next section an overview of our modular face analysis system [8]. An example of how the visual and the acoustic data are integrated to improve robustness of speech recognition is described in section 3.

2. Image Analysis

In a first step the whole image is searched for the presence of heads, and their

locations are determined. Then we zoom in on particular facial features to analyze them in more detail. Regardless whether whole heads or individual facial features are being investigated, the image analysis proceeds in the same way, as shown schematically in Figure 1.

The video stream is processed in three separate channels. The first channel takes monochromatic images as input and searches for the presence of certain shapes and textures. In the second channel color segmentation is done and in the third one the motion is estimated based on frame differences.

2.1. Representation of the Data

Each of the channels produces a set of features and combinations of these features are evaluated with classifiers. Since features produced in different channels may have different representations, it is necessary to provide a scheme to compare different representations.

The set of representations used is shown in Figure 2. They include bitmaps cut from the image after bandpass filtering, binary bitmaps, splines marking the outlines of an area, and simple geometric shapes such as a bounding box or a single point marking the position. Often a very simple representation suffices. For example, when a classifier tries to determine whether three features represent two eyes and a mouth, it takes in a first pass only the center of mass of each feature into account and measures their relative positions. In a next step the classifier also looks at the shape of each feature, at which point the outline representation or, if available, the binary bit map is used.

For each of the representations a distance metric is defined to measure similarity between shapes. The distance metrics are defined between identical representations as well as between different ones.

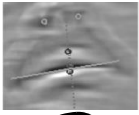


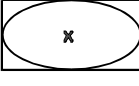
Representations		Distance metric
Filtered bit map (Wavelet filters, Bandpass filters)		Correlation of scaled and rotated bit maps; Profiles
Binary bit map (blob)		Overlap of bit map
Outline, Moments		Overlap Difference of moments
Bounding box Bounding ellipse. Position		Bounding box distance Overlap Euclidian distance
Color information	Histograms	Number of clusters Centers of clusters

Figure 2: Various representations of facial features. For each representation a distance metric is defined to compare features.

2.2. Shape and texture analysis

The first of the system's channels does a shape analysis on monochromatic images. Facial features exhibit intensity variations and, hence, their appearance can be emphasized by selecting an appropriate band of spatial frequencies. Once filtered for spatial frequencies, the image is then 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.

After the filtering operations, the image is thresholded with an adaptive thresholding technique. This produces a binary bit map where areas of interest are marked that can be identified with a connected component analysis. In this way 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 Figure 3.

2.3 Color analysis

Color segmentation is an efficient tool for identifying facial areas. In fact, several studies have been published, where color alone was the feature used for identifying the area of a face. This tends to work well

for conditions where lighting is well controlled. However, color distributions depend strongly on such conditions as lighting and camera characteristics. Therefore, color calibrations can usually not be transferred from one setup to the next. We use color only in combination with the other channels, where we can first calibrate the color space.

For finding a whole face or a facial feature, the color space is clustered with a leader clustering algorithm, where one or two cluster centers are initialized to skin



Figure 3: Examples of processing steps to identify facial features. The top row shows two steps of the shape analysis - left: after bandpass filtering; right: after thresholding. The bottom left image shows the result of the color segmentation. The bottom right displays the result.

colors of a part of the face, identified by the shape analysis. As color space we use a hue, saturation, and luminance (HSL) representation. After skin colors have been identified, the image is thresholded in order to locate the area of the face. Figure 3 shows an example where the face area is identified with the color segmentation. In Figure 6 the mouth area is analyzed with this technique.

2.4. 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 between frames. The resulting difference image is low-pass filtered and thresholded. In the thresholded image areas of moving objects are marked as blobs of connected pixels.

2.5. Combining information

Each of the channels of analysis produce shapes in one or several of the representations shown in Figure 2. Individual shapes carry little information and may mark elements that have nothing to do with a face. Only the combinations of several such shapes are useful to decide whether a face is present. Combining shapes is done with an 'n-gram' search. First each shape is analyzed individually. Those totally out of range in size or aspect ratio can definitely not represent a facial feature and are discarded. Then each of the remaining shapes is labeled with the facial features it might represent, e.g. eye, nostril, head edge, etc.

Next, combinations of two shapes are tested, whether they can represent a pair of facial features, for example an eye pair, eyebrows, or an eye plus a mouth. In the next step triplets are evaluated, etc. In each of these steps the geometric arrangements of

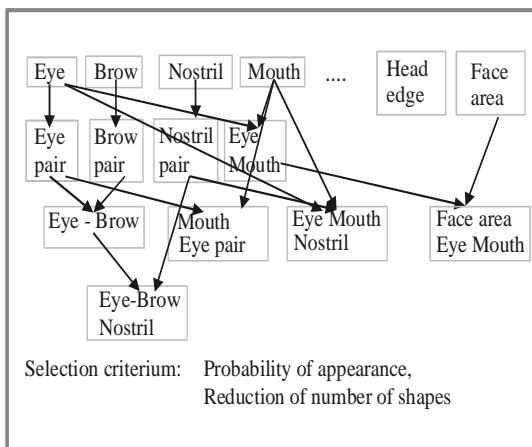


Figure 4: Schematic overview of the n-gram search used to classify combinations of shapes to make a decision whether a face is present in the image. The analysis starts at the top with individual shapes, then evaluates bigrams of shapes, triplegrams, etc.

the features are evaluated with small classifiers that take as their inputs the sizes, ratios of distances, and orientations of the shapes. If information about the reliability of the different channels of analysis is available, the features are multiplied with a weight factor before entering into the classifiers. This n-gram search is shown schematically in Figure 4.

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 by eliminating components with low scores in each step of the search, the computation can be kept fast. In fact, the time for the n-gram search is small compared to that of the band-pass filtering and the shape filtering.

2.6. Models

The classifications, whether combinations of shapes represent a face or not, are done by comparing them with head models. The models are created by measuring the positions of the facial features in sample images and cutting out representative parts. Each of the facial features of a model is stored in various representations. The models are generated from images expected to cover the range of variations that will be encountered in an operating environment. They include faces with and without glasses, with facial hair, and with different skin complexions.

Each model specifies the order in which the n-gram search proceeds and at what point a search can be stopped. For example, for a model of a person wearing a dark moustache the search will start with the moustache and then will try to find two eyes with the proper distance, size, and orientation. If this is found the search stops and returns a high confidence. A light moustache, on the other hand, is not easily differentiated from skin and, hence, is not a reliable feature to detect. In this case the analysis tries to find eyes first and then the mouth, eyebrows, and nostrils. There are multiple search patterns tied to any model. If

the first attempt does not produce a result with high enough confidence, the search starts again with a different combination of shapes.

Figure 5 shows examples of the results from the first pass through the image, where a search is done for whole heads. Several facial features are identified to determine the face orientation. Figures 6 and 7 illustrate results from the analysis of the mouth area. For this example a model with moustache was used to help determine the outlines of the lips.



Figure 5: Examples of identified facial features.

3. Audio-visual speech recognition

The described image analysis system has been tested in audio-visual speech recognition experiments. Key for such experiments is a database large enough to obtain statistically significant results. At the moment there are only a few, relatively small, audio-visual databases available. None of them is large enough, with enough diversity, for speaker independent audio-visual speech recognition. Hence, we had to produce first a suitable database. At the moment it consists of four parts: Two single-speaker databases and two databases with 50 different speakers [9].

3.1. Visual test results

In a speaker-dependent setting, i.e. when the same speaker is used for the training as well as the testing, the reliability for finding



Figure 6: Color segmentation of the mouth area. The left image shows areas with a hue of the lips, the right image indicates hue of the moustache.



Figure 7: Result of the analysis, after combining the shapes from Figure 5 and 6. For a comparison, the right image shows the original.

the lip outlines is very high. After training the system on 30 frames of a person, the lips were located correctly in more than 99% of the over 27,000 frames in the test set. How precisely the lip outlines are marked is difficult to quantify, but an indication of the quality can be derived from the lip reading results. On a data base of 300 five-digit strings the best results showed 93% word accuracy, which is the highest reported for this task so far. This indicates that the accuracy of the mouth location is good.

For an estimate of the accuracy of the lip shapes in the speaker independent case, the system is trained with a randomly selected set of ten speakers and then tested on the other 40 speakers. On average for 87% of the speakers in the test set the mouth location is determined properly in more than 95% of the frames.

3.2. Audio-visual test results

For audio-visual speech recognition the visual data have to be combined with acoustic information. We do this with an early feature integration scheme. This means that feature vectors from the audio and the

visual channels are combined, and this combined vector is then used for the recognition in a two-stream Hidden Markov Model (HMM) classifier.

The two streams of analysis produce feature vectors with different reliabilities. Therefore, for high recognition accuracy this has to be taken into account. In the HMM the two stream observation logarithmic probabilities are combined with different weights. These weights are determined with a training procedure from the statistics of the data [10].

The audio-visual speech recognition has been tested so far on the two single speaker databases. Figure 8 shows some results of the recognition accuracy. For a high signal-to-noise ratio in the acoustic signal the addition of visual information does not improve the recognition accuracy. However, under conditions with substantial noise, the effect of the visual information is quite dramatic. For example, at 10dB SNR, the acoustic signal delivers an accuracy of less than 40%, the visual signal alone 83%, and the combined signal over 95% accuracy.

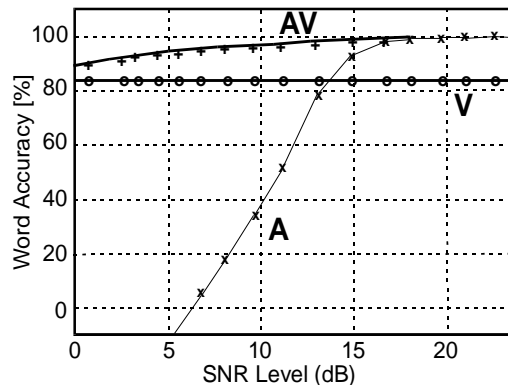


Figure 8: Word accuracy of audio-visual speech recognition for different levels of signal-to noise in the acoustic signal. A: acoustic signal only, V: visual signal only; AV; audio-visual signal. Task: five-tuple digit sequences, single-speaker.

4. Conclusion

Measuring the shapes of facial features is a challenging task, taxing the limits of today's recognition systems. By combining multiple streams of analysis, namely shape, color, and motion, we achieved a high precision in

measuring the shapes of facial features. The good quality of these measurements was confirmed with lip reading experiments, which provided high recognition accuracy. By integrating audio and visual feature vectors with a novel weighting scheme, we demonstrated highly robust audio-visual speech recognition.

References

- [1] **Proc. Int. Conf. Automatic Face and Gesture Recognition**, Killington, IEEE Computer Soc. Press, Los Alamitos, 1996.
- [2] **Speechreading by Humans and Machines**, D.G. Stork and M.E. Hennecke (eds.), Springer, Berlin, 1996.
- [3] **Proc. Audio-Visual Speech Processing**, C. Benoit and R. Campbell (eds.), Rhodes, Greece, 1997.
- [4] M. Sams, V. Surakka, P. Helin, and R. Kättö, **Audiovisual Fusion in Finnish Syllables and Words**, Proc. Audio-Visual Speech Processing: 1997, pp.101-103.
- [5] D. Burnham and S. Keane, **The Japanese McGurk Effect**, Proc. Audio-Visual Speech Processing: 1997, pp.93-96.
- [6] P.L. Silsbee and A.C. Bovik, **Computer Lipreading for Improved Accuracy in Automatic Speech Recognition**, IEEE Trans. Speech and Audio Processing, vol. 4, 1996, pp.337-351.
- [7] J. Luettn, N.A. Thacker, and S.W. Beet, **Active Shape Models for Visual Speech Feature Extraction**, in [2], pp. 383-390.
- [8] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, **Multi-Modal System for Locating Heads and Faces**, in [1], 1996, pp. 88-93.
- [9] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe, **Speaker Independent Audio-Visual Database for Bimodal ASR**, Proc. Audio-Visual Speech Processing: 1997, pp.65-68.
- [10] G. Potamianos, H.P. Graf, **Discriminative Training of HMM Stream Exponents**, Submitted to the Intern. Conf. Acoust. Speech Signal Proc., Seattle, WA, May 1998.