Automatic labelling of facial features

Juan Pablo de la Cruz Gutiérrez[1]

[1] *Infineon Technologies AG, Otto-Hahn Ring 6, München, Email: JuanPablo.delaCruz-Guiterrez@infineon.com*

Introduction

Intelligent interfaces for computers aim at endowing artificial systems with the ability of understanding natural human communication channels like for example speech, facial actions and gestures. For each of those modalities, automatic pattern recognition systems are first to be developed, which are able to identify and represent conveyed information in the form of a collection of measurements. Those systems thus serve as a preprocessing stage for a multimodal platform, which finds commonalities among those modalities.

Essentially, the design of every such system comprises three parts: the collection of a data corpus, the selection of features, and choosing a pattern classification algorithm. The database should contain most significant statistical properties. Sensorial information, like speech or images, show a huge variability, due not only to its intrinsic complex nature, but also to its strong dependence with the environment. A complete database is practically unattainable, and forces researches to limit the scope of their investigations. In order to cope with this limitation, a set of features is pursued, so it constitutes a low-dimensional though robust and complete data representation, in the sense that it accounts for most important data traits. Finally, a pattern classification algorithm able to autonomously discern different classes of elements within the database, together with the proper election of a feature set, overcome to some extent the aforementioned difficulties.

In what follows, a description of an architecture aiming at localizing and tracking most important facial features for human communication is given, following the general taxonomy of the problem described above. This system constitutes a first stage for future projects on lipreading and mimics recognition.

Database

The data corpus used in this task is a collection of images sampled from videos from the CUAVE [3] database. It is a speaker independent audiovisual database of both connected and continuous digit strings totaling over 7000 utterances. Speakers show a wide variety in their appearance: some wear glasses, other beards, as well as different hair styles. We extracted 20 images from each speaker, showing different poses. Sampling was performed for different facial features, namely, nose tip, lips, nostrils and eyes, for these give us relevant information about gaze, pose, facial articulation. Figure 1 shows an example.

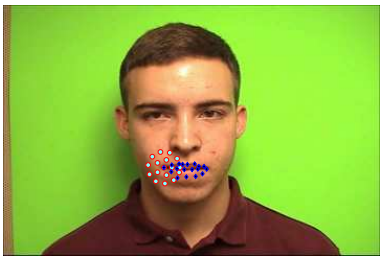

Figure 1: Manual labelling of data. Black diamonds show landmark positions set by hand. Pale blue circles show the distribution of Gabor wavelets around one of the landmarks.

Features

Gabor wavelets are extensively used for object recognition task, for they offer an optimal tradeoff of conjoint resolution in both, the spatial and frequency domains. The extraction of information in terms of undulatory primitives gives us information about structural properties of the surface under analysis, while resolution of positional information indicates the location of the object, at a given scale. Those properties make them a very helpful tool for multi-scale object recognition and identification. They have already been successfully used for a number of different visual object recognition tasks, like face recognition [2], handwritten character recognition and subject identification. At each landmark we

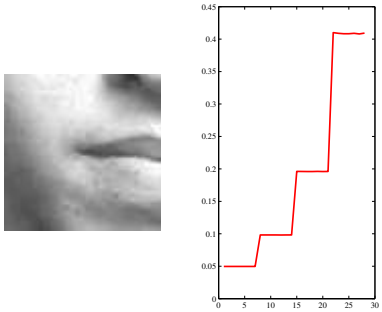

Figure 2: Image region under analysis and plot of absolute value of coefficients of wavelet jet projection. These are plotted sequentially, i.e. coefficients are ordered after their frequencies which explains the plateaus present in the plot. Coefficients show an exponential scaling with frequency, which ought to the election of frequencies and the fact that the energy picked up by Gabor kernels grows with $\|k\|^2$.

calculate the projection of the image onto a Gabor jet (see figure 2) defined by,

$$\psi_{\vec{k}}(\vec{x}) = \frac{\|\vec{k}\|^2}{\sigma^2} \exp\left(\frac{\|\vec{k}\|^2\|\vec{x}\|^2}{2\sigma^2}\right) [\exp(i\vec{k}\vec{x}) - \exp(-\sigma/2)] \quad (1)$$

where

$$\vec{k}_{\nu\mu} = \begin{pmatrix} f_\nu cos\phi_\mu \\ f_\nu sin\phi_\mu \end{pmatrix} \quad (2)$$

is the wave vector and σ determines the ratio of window width to wavelength. Local description of the image is generated at four logarithmically spaced frequency levels ($\nu = 0, \ldots, 4$) and seven orientations ($\mu = 0, \ldots, 6$),

$$f_\nu = 2^{-\frac{\nu+2}{\nu}}\pi, \ \phi_\mu = \mu\frac{\pi}{7} \quad (3)$$

According to those values, the value for the parameter σ was chosen to be two times bigger than the limit imposed by the Nyquist's theorem, so that distortion is small.

Finally, we take the absolute value of projection coefficients. It is commonly believed that most part of image information is contained in the phase information. Shams and von der Malsburg [4] have shown however, that such representation of images is rich enough for visual recognition tasks. Furthermore, the magnitude response of Gabor wavelets is robust against changes in illumination, background, and small distortions.

System overview

Self-organizing maps (SOM) [1], which can be understood as generalization of the vector quantization algorithm, is a useful tool for the visualization of high dimensional signals. It is a neural network model that transforms nonlinear statistical relationships between high-dimensional data into into simple geometrical relationships, so that they can be readily visualized in a display. One of the most interesting traits of this model is that this mapping between feature space and the space of positions of codebook vectors is ordered. In other words, samples close in the feature space, in terms of some metric, are represented by neighboring codebook vectors.

Input data vector consist of wavelet coefficients calculated at every landmark and a set or surround points, in order to capture the local structure of the facial features of interest. On our case, it amounts to 532 coefficients for each point. In order to reduce the dimensionality of the data, PCA transformation was applied, and components account for 95% of the total energy were kept, leading to 12 PCA components. Finally, resulting components of feature vectors were normalized to unit variance. Simulations were carried out by means of the MatLab toolbox developed at the University of Helsinky [6], which implements the basic SOM algorithm and a set of helpful visualization functions. Since we attempt at locating both, faces and facial features, firsts experiments focused on detection of lips. This first trial already showed some of the basic difficulties, and the need for additional preprocessing stages of images.

Results and conclusions

Figure 3 shows an image in which the points where the best founded matches for lips are represented with crosses on top of the subject's face. Results on database show that the algorithm aims at the area around the mouth in most cases, it is sensible to some distracters like beards and faces which show very low contrast (round pale faces under high illumination conditions). As a first attempt to correct for the latter case, we applied illumination gradient correction (substraction of best-fit brightness plane) together with histogram equalization [5] which enhances facial features. Nonetheless, the number of distracters increased considerably.

One of the causes is the lack of a sufficient number of samples. Experiments with the SOM showed that the borders of classes are not well-defined. Besides, there is an important open question which also influences the performance of the system, namely, a criteria for the selection of best value for parameter σ, and hence the rest of wavelets parameters.

Figure 3: Blue crosses show position of best matches for lips.

References

[1] Teuvo Kohonen. *Self-organizing Maps*, volume 30 of *Information Sciences*. Springer, 3rd edition, 2001.

[2] Martin Lades, Jan C. Vorbrüggen, Joachim Buhmann, Jörg Lange, Christoph von der Malsburg, Rolf P. Würtz, and Wolfgang Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–310, March 1993.

[3] E. K. Patterson, S. Gurburz, S. Tufekci, and J. N. Gowdy. Moving-talker, speaker independent feature study and baseline results using the cuave multimodal speech corpus. In *Eurasip 2002*, 2002.

[4] Ladan Shams and Christoph von der Malsburg. The role of complex cells in object recognition. *Vision Research*, 42:2547–2554, 2002.

[5] Kah-Say Sung and Tomaso Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, January 1998.

[6] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. *SOM toolbox for Matlab 5*. Helsinki University of Technology, http://www.cis.hut.fi/projects/somtoolbox/, April 2000.