Image Features Based on Two-dimensional FFT for Gesture Analysis and Recognition

Paul Modler*, Tony Myatt†

* Hochschule für Gestaltung, Germany, pmodler@hfg-karlsruhe.de

† University of York, Great Britain, am12@york.ac.uk

***Abstract* — This paper describes features and the feature extraction processing which were applied for recognising gestures by artificial neural networks. The features were applied for two cases: time series of luminance rate images for hand gestures and time series of pure grey-scale images of the facial mouth region. A focus will be on the presentation and discussion of the application of 2-dimensional Fourier transformed images both for luminance rate feature maps of hand gestures and for greyscale images of the facial mouth region. Appearance-based features in this context are understood as features based on whole images, which perform well for highly articulated objects. The described approach was used based on our assumption that highly articulated objects are of great interest for musical applications.**

I. Spatial APPearance Based Features

A. *Feature Maps of Luminance Rate (Optical Flow Grade Zero)*

The visual energy of two consecutive video frames may be used as a feature map for the recognition of a visual time series such as hand gestures. This approach is understood as luminance rate. Different denotations are used for the luminance rate feature, such as optical flow of grade zero [23], difference image or visual energy. This has been used in several approaches both in scientific-technical and musical/artistic environments [31], [32].

Advantages of this approach are:

- Masking the (stationary) background
- Robustness against variations of lighting such as:
  - Intensity
  - Contrast
  - Light Temperature
- There are some suggestion that the approach is close to biological mechanisms (i.e. the high attention to motion in the visual cortex)
- Fast to compute

The drawbacks are:

- Motion in the background scene detracts from the observed object
- Related motions detract from the target (e.g. a motion of the hand is often combined with a motion of the whole arm)
- Luminance rate is zero for still objects
- There is a lower significance for slow moving objects

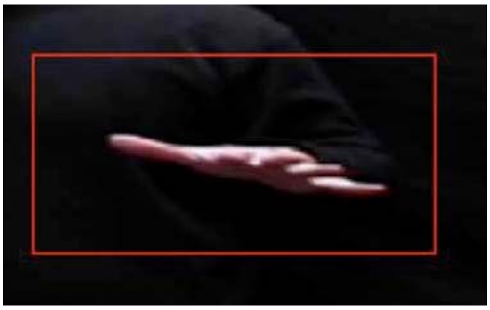
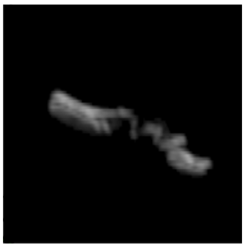


Fig. 1: Tracked left hand with attention rectangle (image size 640x240)

 Fig. 2: Luminance rate feature for attention rectangle of Fig. 1 (frame size 160x80)	 Fig. 3: Luminance rate feature for attention rectangle of Fig. 1 (down-sample dimensions 32x32)

A modified Cam-Shift algorithm may be used to focus the attention on the relevant part of the video stream i.e. the hand as shown in Fig. 1. For the context of this work this region of attention is named the Attention Rectangle (AR) or Region of Interest (ROI) as used in OpenCV [23]. The luminance rate image may be computed from the difference of two consecutive grey scale images of the AR according to Eq. 1.

$$LumaRate(x_{AR}, y_{AR}, t) = Luma(x_{AR}, y_{AR}, t) - Luma(x_{AR}, y_{AR}, t-1)$$

Eq. 1: Computation of the Luminance Rate,
xAR,yAR: spatial coordinates of the Attention-Rectangle

B. *Feature Maps of pure Grey Scale Images (Mouth Gestures)*

Images of mouth gestures were used to investigate the recognition of static poses by recognition algorithms such as artificial neural networks. This was motivated by the prospect that, in addition to using energy as an intuitive control for musical parameters, position and force were both associable with stable gesture states. Furthermore gestures of the facial mouth region differ in the form of the gesture object (the mouth region) and the type of motion, speed, background and position.

Facial features of the mouth region are expected to have

- Fewer rotational variation in the image frame (head is kept up)
- Fewer background variations (faces are similar)
- Gesture variations are more of a textural nature (with similar background, the face)

An initial estimate identified four feature types to be considered for use with a gesture recognition algorithm as shown in Table 1

	Pro	**Con**
Extracted Colour Feature	Clear feature regions (white or grey on surrounding black), appearance similar to luminance rate of hand gestures	Unstable border regions unstable position unnatural feature: lips have to be coloured
Grey Scale Image	Fast, reduced complexity, poses possible	Problems of variations of light intensities, shading,
Edge Extraction	Fast computation, Increased robustness against variations of light, background, colour	Tolerant to shading View relevant pixels
Luminance Rate	Approach similar to hand gestures	Low feature intensity due to slow motion of the gestures and more textural motions

Table 1: Considerations on features for mouth gestures

Assuming a coloured marker for tracking the mouth region, a colour criterion may be used as shown in Fig. 4. The resulting images of the AR are similar to images produced by the luminance rate: a blob or cluster with a high intensity in the centre region of the AR surrounded by pixels with a low or zero intensity as shown in Fig. 6 and in Fig. 7. Alternatively a grey scale image may be used to provide the whole image structure of the AR as shown in Fig. 5. A third alternative is to extract edges of the image for example by using a Sobel operator as shown in Fig. 9. The fourth choice, the luminance rate as shown in Fig. 8, was ruled out due to its reduced intensity.

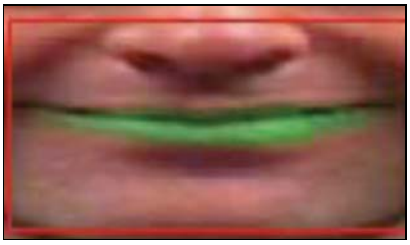


Fig. 4: Video Image for gesture grinning (640x240 pixels) AR tracked with green marker

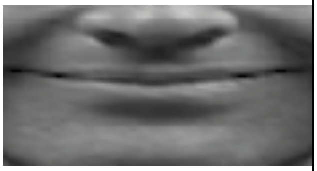


Fig. 5: Pure greyscale image of AR (120x60) of Fig. 4

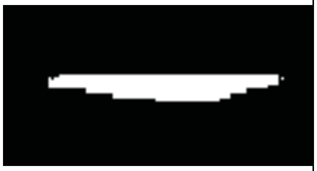


Fig. 6: Green colour feature of AR of Fig. 5

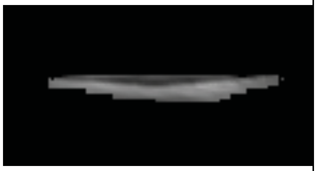


Fig. 7: AR of Fig. 4 keyed by green colour

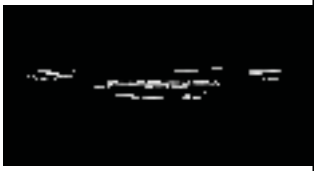


Fig. 8: Luminance Rate feature for AR of Fig. 4

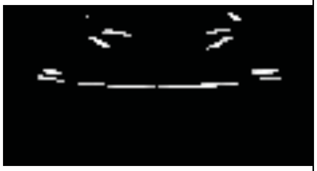


Fig. 9: Edge feature by Sobel operator threshold 0.77 of AR of Fig. 4

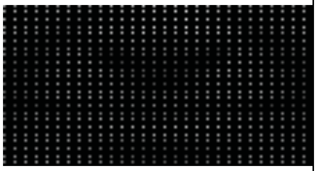


Fig. 10: Pure greyscale image of AR of Fig. 4, downsampled to 30x30 pixels

To provide an AR for the mouth region, similar to the approach for hand gestures, a Cam-Shift algorithm was used. Due to the relatively low visual energy of mouth gestures, as well as the aim to investigate static poses, a distinct colour for tracking the mouth and providing the coordinates of the AR was chosen.

The colours were selected to ensure that the marked grey scale images should, as far as possible, have similar properties to unmarked grey scale images. This was based on the intention of using the marked images for approaches, which were not based on colour markers.

II. APPLICATION OF APPEARANCE BASED FEATURES FROM SPATIAL FOURIER TRANSFORMED IMAGES

Two dimensional spatial frequency transformations are widely used in image processing. The common JPEG image compression is based on a spatial Discrete Cosine Transformation (DCT) also used in MPEG-2 coding of moving images [18]. Wavelet transformations for images offer high compression rates [26].

In general, transformations provide the advantage of describing data in a different space, which has distinct properties. For example a one-dimensional temporal Fourier transformation (Eq. 2 and Eq. 5) converts an audio signal into a magnitude and phase spectrum (Eq. 3 and Eq. 4) providing a means of analysis in addition to applications, which are not possible in the time-domain [5]. A complex harmonic oscillation convolved with the time-domain signal to achieve the Fourier transformation splits the original audio signal into an intensity dependent real part and a time dependent imaginary part.

$$F_{(\omega)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f_{(t)} e^{-j\omega t} dt$$

Eq. 2: Continuous one-dimensional Fourier transformation

$$M(\omega) = \sqrt{\mathrm{Im}\,ag^2(F_{(\omega)}) + \mathrm{Re}\,al^2(F_{(\omega)})}$$

Eq. 3: Magnitude of 1d-transformed Fourier transformed

$$P(\omega) = \tan^{-1}\left(\frac{\mathrm{Im}\,ag(F_{(\omega)})}{\mathrm{Re}\,al(F_{(\omega)})}\right)$$

Eq. 4: Phase of 1d-transformed Fourier transformed

$$F_k = \sum_{n=1}^{N} x_{xn} e^{-j2\pi\frac{kn}{N}}$$

Eq. 5: Discrete one-dimensional Fourier transformation

In the temporal frequency domain (audio) this property appears in form of a magnitude spectrum where single spectral lines contain no information about their temporal position and the phase spectrum where the temporal position (the phase) of the spectral line is given but with no information about its intensity. An interesting application of this is the Mammut software [22].

Based on this split representation of the original signal, further analysis and processing may be undertaken such as data reduction through disregarding irrelevant components or filtering by multiplication of a desired magnitude spectrum.

Analogous to the one-dimensional case, the two-dimensional Fourier transformation is formulated according to

Eq. 6 and its magnitude and phase spectra according to Eq. 7 and Eq. 8.

$$F(u,v) = \iint f(x,y) e^{-j2\pi(ux+vy)} dxdy$$

Eq. 6: Two-dimensional continuous Fourier transformation

$$M(u,v) = \sqrt{\mathrm{Im}\,ag^2(F_{(u,v)}) + \mathrm{Re}\,al^2(F_{(u,v)})}$$

Eq. 7: Magnitude M of 2d-Fourier transformed

$$P(u,v) = \tan^{-1}\left(\frac{\mathrm{Im}\,ag(F_{(u,v)})}{\mathrm{Re}\,al(F_{(u,v)})}\right)$$

Eq. 8: Phase P of 2d-Fourier transformed

As for the one-dimensional Fourier transformation, fast versions of the discrete formulation are available [4], [12]. The open source package “Fastest Fourier Transformation in the West” [12] provides computation of multidimensional FFTs at low CPU costs and unrestricted dimension sizes such as dimensions different from powers of two.

Two properties of the two-dimensional spatial transformation are of main interest:

- The separation of an image in an intensity based magnitude spectrum and a position based phase spectrum should increase the position independence of a later applied recognition [6]
- A possible reduction to relevant portions of the spatial image spectrum

Other transformations such as Wavelet or Gabor transformations may improve other areas of the recognition problem, such as independency of rotation, size and shading. A more complete description of the properties of two-dimensional Fourier transformations may be found in [4].

A. Image Representation based on 2D-FFT

A Fourier transformed image in general is displayed as a magnitude spectrum with interchanged quadrants shifting the low frequencies to the centre of the image and the high frequencies to the borders as shown in Fig. 11 and Fig. 12. As outlined previously, the 2d- Fast Fourier Transformation has the property that positional information in an image is represented in the phase spectrum of the image which can be used to achieve an image representation where the position constraints are less important leading to a gradual position independence.

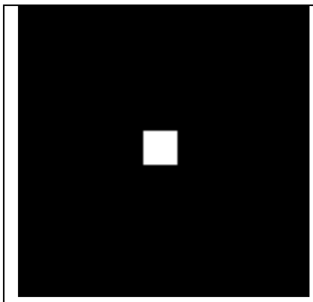
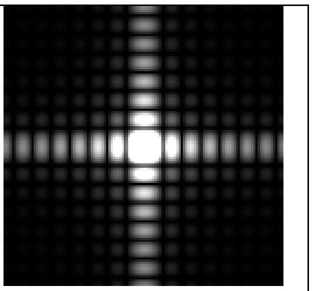
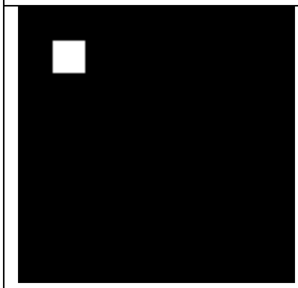
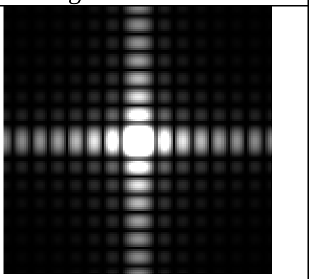
 Fig. 11: Rectangle in image	 Fig. 12: Transformed of rectangle
 Fig. 13: Shifted Rectangle	 Fig. 14: Transformed of shifted rectangle

Image compression is achievable by omitting higher frequencies of the spatial image spectrum [4]. This is an alternative to the pixel-wise smoothing and down sampling algorithm used in the spatial domain The overall image of the object (hand, mouth) is still preserved for the FFT compression whereas in the down sampling version the image tends to turn into a noisy disturbed region. Drawbacks of this are increased computation operations, such as:

- Computation of the 2d-FFT

- Computation of the magnitude
- Data handling

Due to the real value input of a grey-scale image, the resulting magnitude spectrum of an FFT is symmetric which was incorporated in the current approach by only using one half of the magnitude spectrum. Images of transformed images from my implementation show only one half of the spectrum e.g. Fig. 17.

B. 2D-FFT Window

A windowing image, which smoothes the edges of the spatial FFT area, reduces the influence of shading and translational variations on the spectrum of the image. The windowing image is multiplied with the AR prior to the FFT. A similar approach is described by [14]. As a first approach I used windows with linear transitions as shown in Fig. 15. The width of the transition regions is 15 pixels for horizontal and vertical borders. Other window types such as such as a Hanning or Hamming window may be considered.

Fig. 15: FFT-window (120x60), with linear transitions (15 pixels)

An increased width of the transition borders of the window will increase the shift invariance of the transformed greyscale image as long as the image contains frequencies in the border regions. The windowing may be seen as impressing a single blob shape onto a greyscale image. For blob like images such as luminance rate images the windowing will have a less significant role.

C. Implication of Intensity Normalisation of Images to their Magnitude Map

To deal with overall varying contrast and light conditions a normalization of the mean image intensity was applied (Eq. 9 and Eq. 10). A resulting pumping effect of the intensity of the image was accepted. A temporal smoothing of the image mean $\bar{I}_t$ may be considered to reduce this effect. Fig. 21 shows the Fourier magnitude map of the normalised Fig. 20.

$$\bar{I}_t = \frac{\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1} I_t(x,y)}{X*Y}$$

Eq. 9: Mean Intensity $\bar{I}_t$ of an Image with dimensions X, Y and pixel intensity I(x,y,t)

$$In_t(x,y) = I(x,y)_t - \bar{I}_{t-1}$$

Eq. 10: Normalised intensity In(x,y,t) of a pixel

Due to the subtraction of $\bar{I}_t$, resulting images may have a negative value. In fact Eq. 10 may be seen as a DC offset removal, shifting the zero plane into the middle of the spatial waveforms. For a following Fourier transformation this has no disturbing effects, but to display the resulting values they have to be shifted again into the positive domain. This has to be considered when analysing normalised images such as Fig. 20. From the above images it is not obvious whether an intensity normalised image sequence is more robust against overall intensity changes or shading if the images are high-pass filtered. For this a pattern-set was created in which the low frequency row and column was suppressed, but without an intensity normalisation.

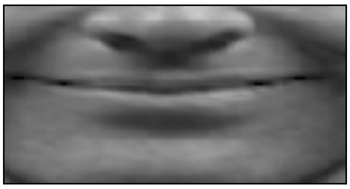
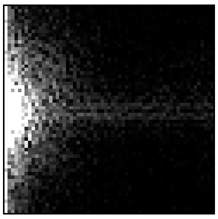
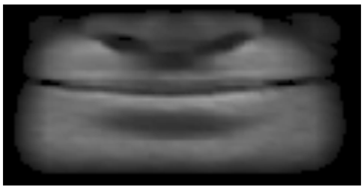
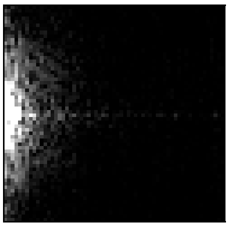
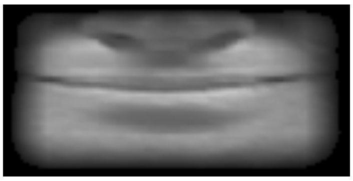
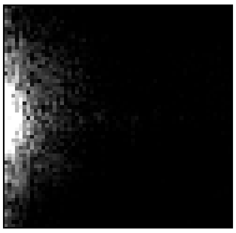
Image (120x60 points)	Transformed (61x60 points)
 Fig. 16: Pure greyscale image of AR of Fig. 4	 Fig. 17: FFT of Fig. 16
 Fig. 18: Windowed image	 Fig. 19: FFT of Fig. 18
 Fig. 20: Windowed and removed DC offset	 Fig. 21: FFT of Fig. 20

Table 2: Transformed images through applied processing stages

D. Shift Invariance through Omitting Spatial Phase Spectrum

To exploit the shift invariance of the magnitude spectrum the 2d phase spectrum of the image was omitted.

E. Data Reduction by Frequency Truncation: 2D High-Cut Filter

To achieve a desired data reduction the magnitude of a transformed image was truncated by pruning high frequency areas as described by [4]. There the author claims a high image reduction by punching out the centre portion of the transformed image. This procedure can be seen to omit all high frequent information of the image, which is similar to spatial smoothing of the original image. It is assumed that high frequency areas carry less important information and so can be omitted while preserving the major characteristics of the image. The amount of data reduction required is primarily relevant to the size of the training data and the time required to train the neural network. Since this work aims to achieve a real time application, CPU costs are shifted from the computation size of the neural network to the computation of the 2D-FFT.

For the case of hand gestures, experiments with different high cut filters with varying cutoff-areas have been undertaken, showing that a great amount of data reduction is achievable while preserving the information necessary for recognition of objects by an artificial neural network.

Dimensions u,v	**Number of Feature points**	**Reduction relative to original image size (160x80=12800 pixels)**
160x80	12800	0%
32x32	1024	92,0 %
16x16	256	98,0 %
8x8	64	99.5 %
4x4	8	99.9375 %

Table 3: Size of truncated magnitude spectrum and achieved data reduction

F. *Spatial Resolution and Aliasing*

Further optimisation may be considered such as applying a 2d-FFT with a reduced number of points. In this case the image has to be filtered before the transformation to avoid aliasing in the lower frequency areas. In the current approach aliasing is avoided by transforming the image of the AR using the original image resolution and then truncating the high frequency areas.

G. *Reduction of Variations caused by Shading: Low Cut Filtering*

Variations of gradual shading may be simulated through the multiplication of an intensity ramp matrix to the source image. This can be seen as the superposition of a low frequency wave with a wavelength of a quarter of the relevant image size and a phase according to the direction of the ramp. A whole window with a shading function from light (left) to black (right) has a magnitude of a straight horizontal line. It may be interpreted that the shading brings out the windowing artefacts of the FFT but only in the x direction due to the sharp edges at the lower and upper border of the window from dark to light.

To generalize the extracted spatial frequency features concerning shading variations the low frequency areas of the transformed images were suppressed, which is equivalent to a low cut filtering.

Above assumptions are confirmed using a Max/Jitter patch to resynthesise mean intensity normalized images. The visual results of synthetically shaded image were compared to determine approaches to how to treat low frequency components such as dropping the whole lowest frequency row and column (x = 0, y = 0 to N; and x = 0 to N, y = 0) or dropping of only 2 low frequencies.((x=0, y=0 and x = N-1, y = 0). Fig. 23: which is not shaded and not normalised differs only slightly from Fig. 27 (not shaded, normalised) due to the zeroing of the low frequencies where most of this information is coded.

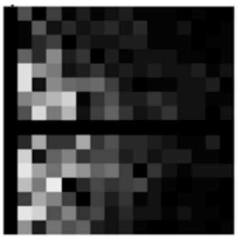
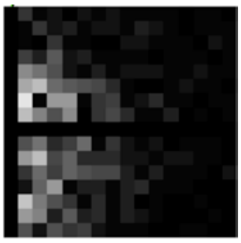
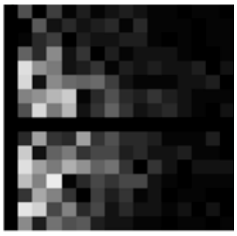
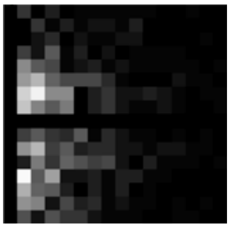
Image (120x60 pixels)	Transformed and truncated high frequencies and zeroed low frequencies (16x16 points)
 Fig. 22: Not shaded, not normalised	 Fig. 23:
 Fig. 24: Shaded (darker left side) not normalised	 Fig. 25
 Fig. 26: Not shaded normalised	 Fig. 27
 Fig. 28: Shaded normalised	 Fig. 29

Table 4: Influence of shading and normalisation on the magnitude spectrum

H. *Grouping of Frequency Areas*

To further reduce data while preserving relevant content as far as possible, a grouping of frequency groups into frequency bins is possible. Comparable methods are used in frequency based coding or visualisation of audio signals. The size and number of bins were heuristic and led by the assumption that the relevance of single frequencies is more important for lower frequencies than for higher, this assumption is analogue to human audio perception which is closer to a logarithmic scale than the linear scale implied by the Fast Fourier Transformation.

I. Logarithmic Coding of Frequency Values

In the image space a pixel is assumed to be an 8 bit value for one colour channel, which holds for a wide range of current consumer and low-end professional video devices. This results in a minimal resolution of 1/255 equivalent to 0.003921568627. To compress the data value range in the frequency domain, a logarithmic scaling of the values was used offering the possibility to use the negative values for the input neurons of the neural network. The logarithmic scaling can also be seen to stretch the value range between 0.0 and 1.0 (from negative infinity to 0) and to compress the value range above 1. To avoid exceptions (NAN) during the computation of the logarithm frequency values with value 0.0 an offset is added to all frequency values F(u,v).

$$F_{\log}(u,v) = \log_{10}(F(u,v) + a) \qquad a = 1.0, a << 1.0$$

Eq. 11: Log frequency

The offset a=1.0 eliminates the negative log values, whereas a very small offset (a<<1.0) preserves negative values, but introduces large negative frequency values.

J. Variations of Image Contrast

Contrast variations have to be considered for different camera and lighting setups. Contrast variations relevant for the recognition are variations in the AR, resulting in the need to measure the ARs contrast for an image sequence.

1) Contrast Modification for Pure Grey Scale Images

Modification of contrast of one image was implemented according to Eq. 12 and Eq. 13.

$$E(t_0) = \sum_{y=0}^{Y}\sum_{x=0}^{X} P(x,y,t_0)$$

Eq. 12: Mean Intensity E(t0) at time t0 of an image
(X=120, Y=60)

$$P_c(x,y,t_0) = \big(P(x,y,t_0) - E(t_0\big) * C_f + E(t_0)$$

Eq. 13: Contrast modification of Pixel P(x,y,t$_0$) to P$_c$(x,y,t$_0$),
Cf contrast modification factor: 1/Cf ~ Pmax-Pmin

2) Contrast Modification for Luminance Rate Images

For luminance rate images generated from difference images a contrast transformation can be derived directly from he subtraction of the mean image, assuming small and slow variations of the image content.

$$P_{LRc}(x,y,t_0) = P_c(x,y,t_0) - P_c(x,y,t_{-1})$$

Eq. 14: Contrast modification for difference images

$$E(t_0) - E(t_{-1}) \lhd\lhd 1.0 \Rightarrow E(t_0) \approx E(t_{-1})$$

Eq. 15: Slow changing image mean E(t)

$$C_f(t_0) \approx C_f(t_{-1})$$

Eq. 16: Resulting similar contrast factor C$_f$

For slow temporal changes of image content contrast variations of luminance rate images are small (Eq. 16). This leads to a more robust recognition of luminance rate feature patterns distorted with contrast variations compared to greyscale patterns
The application of an optimised normalisation and contrast adoption algorithm and the reduction of the resilience to shadow casts should improve the detection performance. Extending the training pattern set by using images with variations of the shadow cast and aspect should reduce the effect of shadow cast significantly.

K. Appearance-Based 2D-FFT: Summary of Algorithm Steps

Following list summarises the processing steps to generate a pattern frame which was fed to the TDNN:

1) Tracking of the facial mouth region
2) Conversion of the image of the attention rectangle to a greyscale intensity image
3) Normalisation (DC-Offset removal, contrast balance)
4) Multiplication of 2d-window (shift invariance)
5) 2d-FFT
6) Computation of the 2d-magnitude
7) Low-cut filtering (shading)
8) High-cut filtering (data reduction and smoothing)

L. Results

To verify the functionality of the process Time Delay Neural Networks were trained with a pattern set generated from grey-scale images of 120x60 pixels of the mouth region, tracked by green marked lips. The high frequency areas of the magnitude spectrum of the transformed images were truncated to 16x16 points. The two lowest frequency points were zeroed for all patterns. The network was tested using a movie with recorded test gestures. The overall impression of the resulting recognition was very good and is discussed below:

- *Data Reduction:* The truncated 16x16 sized feature maps were able to provide enough content for the recognition algorithm to recognise the gestures of the set which is equivalent to data reduction of 96.55%
- *Transversal shift invariance:* In contrast to the untransformed processing the windowed and phase-less transformed feature maps provided shift robustness of about 7 to 10 pixels and more,

depending on the width of the window borders and the content of the image.

- *Intensity variations:* The intensity normalised magnitude spectrum provided a large generalisation of synthetic (movie brightness) variations of light intensity such as almost dark to very light.
- *Gradual shading:* The low cut filter applied though zeroing two frequency points provided robustness against synthetic shading as far as the shading did not obscure relevant parts of the image. For example, shading the right side of the mouth in a right side smirking gesture prevented the network recognition.
- *Lighting contrast:* Although the intensity normalisation and the intensity variations of the training patterns increased robustness against contrast variations, the recognition was resilient to contrast variations
- *Shadows:* The network was highly resilient to variations in the casting of shadows caused, for example, by a different lighting position. An increased size of the low frequency cut area significantly improved the performance of the overall recognition but reduced the ability of the network to distinguish between gestures.
- *Performance and CPU costs:* Although the transformation and related processing increased the cpu costs the Max/Msp implementation of the system performs in real time (25 fps) on a 1.5GHz G4 Powerbook using an IEEE 1394 video input device.
- *Luminance rate:* Due to the increased robustness of the luminance rate against gradual shading, the low frequency cut filtering can be omitted. Furthermore position variations of the trained gestures (gestures recorded at 5 different positions) are assumed to have provided an increased robustness against variations of shadow casts by different light conditions.

III. Conclusions

In this paper the features and feature extraction process were described both for pure grey-scale images and luminance rate images. Based on previous applied processing methods, the approach to using Fourier transformed images for both cases were described in detail. The results showed that the transformation based approach is a valid tool to:

- Reduce the data while retaining relevant content for the gesture recognition
- Increase the shift invariance of the recognition
- Provide features for gesture recognition based on greyscale images
- Provide feature extraction in real time

Problems may be addressed as follows:

- The ANN recognition algorithm showed the tendency to interchange gestures differing mainly in their motion direction
- The recognition algorithm showed an increased sensitivity to variations of shadow casts caused by lighting variations
- The recognition is sensitive to contrast variations

For the integration into an interactive computer music environment appearance based features of Fourier transformed images may be combined successfully with artificial neural networks to recognise gestures of the hand or gestures of the facial mouth region

References

[1] F. Althoff, R. Lindl, L. Walchshäusl , Robust Multimodal Hand- and Head Gesture Recognition for controlling Automotive Infotainment Systems. VDI-Tagung - Der Fahrer im 21. Jahrhundert, Braunschweig, Germany, 21.11.2005

[2] Becker, D.A.: Sensei: A Real-Time Recognition, Feedback and Training System for T'aiChi Gestures. M.I.T. Media Lab Perceptual Computing Group Technical Report No. 426, 1997.

[3] Bishop, M., Neural Networks for Pattern Recognition, Oxford University Press, 1995

[4] Bow, S.T.: Pattern Recognition and Image Preprocessing. Marcel Dekker, Inc. 2002.

[5] Bracewell, R. N., The Fourier Transform & Its Applications, McGraw-Hill Science/Engineering/Math, 3 edition, 1999

[6] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. Proc. ICASSP, 1993. Minneapolis.

[7] R. Brunelli, T. Poggio, Face Recognition: Features versus Templates, IEEE Trans. on PAMI, Vol. 15, No. 10, pp. 1042-1052, Oct. 1993

[8] L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick, A. Pentland, "Invariant features for 3-D gesture recognition," fg, p. 157, 2nd International Conference on Automatic Face and Gesture Recognition (FG '96), 1996.

[9] Cassell, J.: A Framework For Gesture Generation And Interpretation. In Cipolla, R. and Pentland, A. (eds.), Computer Vision in Human-Machine Interaction, pp. 191-215. New York: Cambridge University Press. 1998.

[10] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," Proc. Third IEEE Conference on Face and Gesture Recognition, Nara, Japan, April 1998

[11] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis "Exemplar-based Tracking and Recognition of Arm Gestures" 3rd International Symposium on Image and Signal Processing and Analysis (ISPA 2003) Rome, Italy, September 18-20, 2003

[12] (FFTW, 2006) , http://www.fftw.org/ (checked 13.11.2006)

[13] Kanade, T, Computer recognition of human faces. Birkhauser, Basel, Switzerland, and Stuttgart, Germany, 1973

[14] Mathias Kolsch, Matthew Turk, "Robust Hand Detection," fgr, p. 614, Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004

[15] (Lee, Xu, 1996)

[16] Christopher Lee and Yangsheng Xu, Interactive Learning of Gestures for Human Robot Interfaces, IEEE Int. Conf. on Robotics and Automation, pp 29822987, 1996

[17] Modler, Paul, Zannos, loannis, Emotional Aspects of Recognition by Neural Networks, using dedicated Input Devices, in Antonio Camurri (cd.) Proc. Of KANSEI The Technology of Emotion, AIMI International Workshop, Univcrsita Genova, Genova 1997

[18] The MPEG Homepage, http://www.chiariglione.org/mpeg/ (

[19] C. Neustaedter, "An Evaluation of Optical Flow using Lucas and Kanade's Algorithm," 2002.

[20] Nickel, K., Stiefelhagen, R.: 3D-Tracking of Heads and Hands for Pointing Gesture Recognition in a Human- Robot Interaction Scenario, Sixth Int. Conf. On Face and Gesture Recognition, May 2004, Seoul, Korea

[21] Kai Nickel, Tobias Gehrig, Hazim Kemal Ekenel, John McDonough, Rainer Stiefelhagen An Audio-visual Particle Filter for Speaker Tracking on the CLEAR06 Evaluation Dataset

CLEAR Evaluation Workshop, CLEAR Evaluation Workshop 2006, Southampton, UK, 2006-04

[22] http://www.notam02.no/notam02/prod-prg-mammuthelp.html (checked 11.11.2006)

[23] Open Source Computer Visions Library, Reference Manual, Intel Corporation 1999-2001

[24] Park, T. H., P. Cook 2005. "Nearest Error Centroid Clustering for Radial/Elliptical Basis Function Neural Networks in Timbre Classification". Proceedings of the 2005 ICMC, Barcelona, Spain

[25] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. pages 228-233, 1998

[26] Salomon, D. Data Compression. The Complete Reference, Springer, Berlin, 2004

[27] Mingli Song, Jiajun Bu, Chun Chen, Nan Li, "Audio-Visual Based Emotion Recognition A New Approach," cvpr, pp. 1020-1025, 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04) - Volume 2, 2004

[28] Andreas Tewes, Rolf P. Würtz, and Christoph von der Malsburg. A flexible object model for recognising and synthesising facial expressions. In Takeo Kanade, Nalini Ratha, and Anil Jain(eds.): Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication, pp. 81-90. Springer LNCS, 2005

[29] Matthew Turk, Gesture Recognition, in, http://vehand.engr.ucf.edu/handbook/ Microsoft Research 2006

[30] Valette, S., Magnin I., Prost R., Mesh-based video objects tracking combining motion and luminance discontinuities criteria, Signal Processing archiv, Volume 84 , Issue 7 (July 2004) Pages: 1213 – 1224, 2004

[31] KAMEDA Yoshinari, MINOH Michihiko, A Human Motion Tracking Method Using Double Difference Technique", Proceedings of Conference of Kansai District Union of Electric Related Institutes, S12-3:S64, 1996.

[32] Modler, P. Myatt, A., Saup, M., An Experimental set of Hand Gestures for Expressive Control of Musical Parameters in Realtime, 2003 International Conferences on New Interfaces for Musical Expression Proceedings, McGill University, Montreal, 2003