УДК 621.382

FPGA BASED REAL TIME HAND DETECTION BY MEANS OF COLOUR SEGMENTATION

M. KRIPS, J. VELTEN, A. KUMMERT

Communication Theory
Faculty of Electrical, Information and Media Engineering,
University of Wuppertal,
Rainer-Gruenter-Strasse 21, 42119 Wuppertal, Germany

Submitted 20 November 2003

Colour is an important feature for object classification in digital video images, where shape-based techniques are not practicable. This can be the case, if real-time processing has to be applied to situations, where many candidate shapes can occur. However, colour can be represented in different colour spaces. Which of them is appropriate cannot be fixed in general for all purposes. In this paper, the influence of colour space is illustrated by an FPGA based real time hand detection system. The latter presents a challenging problem, where shape-based approaches are unpromising. This detection system consists of an artificial neural network, which classifies particular single pixels whether they belong to the hand or to the background. Successive use of binary morphological operations improves results.

Keywords: artificial neural network, colour spaces, digital image processing, binary morphological operation, filed programmable gate-array.

Introduction

Digital video image processing is of growing interest, due to decreasing prices for (digital) cameras and powerful computers. Range of applications enlarges rapidly, especially in the field of industrial applications. Video image analysis is applied to tasks, where correct positions or orientations of components must be checked. For most of these tasks a monochrome camera is used and the shape of objects enables classification. If there are too many different shapes for one object, classification time explodes. An example for such a problem is hand detection in digital video images. Here, classification based on shape comparison is inapplicable. Especially, if detection has to be done in real time, other features have to be used. Colour can be such a powerful feature, which can be used by using a colour charge coupling device (CCD) camera. For hand detection, not only a single colour but a set of colours has to be classified as belonging to hand and skin, respectively. Therefore, an adaptive segmentation approach by means of artificial neural networks has been chosen. In this context, different colour space representations have to be investigated concerning their suitability for neural network based classification schemes. This aspect is of major importance for the presented detection system, since classification has to be done based on the information of a single pixel. Information stemming from the neighbourhood of the actual pixel is not taken into account at this first processing step. The advantage of this single-pixel based approach can be seen in the fact that very high processing speed can be obtained, since a continuous data stream is processed. Furthermore, no storage of complete colour images is needed during the classification process.

Classification

Detection of hands in digital video images can be described as a classification of each single input pixel into one of two classes ε_1 and ε_2, where ε_1 denotes the class of hand and skin, respectively, and ε_2 denotes the background. Obviously, this classification is not a linearly separable problem. Therefore, a multilayer feed-forward network is used for classification purpose [1]. The choice of this type of neural network was endorsed by the demand on a reconfigurable hardware implementation by means of an FPGA (Field Programmable Gate-Array). The latter can be easily used since only a few modifications of the algorithms developed on a software platform are required for an implementation by VHDL [2]. The neuron's architecture consists of a linear combiner followed by an activation function. Except for the output neuron, where a logistic function is used, the hyperbolic tangent is applied. Processing time minimization implies the use of a neural network, where number of neurons and layers is as small as possible. If the neural network is realized as a software solution on a computer, both aspects have influence on processing time. However, if it is implemented in hardware, the advantage of parallel processing can be used. In this case, processing time is proportional to the number of layers but does not depend on the number of neurons within a layer. Nevertheless, the number of neurons per layer cannot be increased arbitrarily, since the effort of hardware will increase accordingly. Therefore, the size of neural networks varies between 3/3/1 and 3/3/4/8/4/1, where 3/3/1 denotes a network with three neurons in the input layer, three neurons in the hidden layer and one neuron in the output layer. Please notice, the input layer only provides the inputs and does not require computations. All networks considered consist of three input neurons, as information of a single pixel is given by a tristimulus in all colour spaces considered in the following. For all of them the training of neural networks has been done by a variant of backpropagation, namely Levenberg-Marquardt [3]. For network learning, a training pattern consisting of a subset of an original frame and a binary reference output created manually by means of graphical software was used.

Colour spaces. Since there are many different colour space representations [4–8] only some can be discussed here. In the following, RGB, HSV, YIQ (NTSC), and YC_bC_r (PAL) will be shortly introduced.

RGB. The RGB colour space is the most common one in the field of digital image processing. The acronym RGB stands for Red, Green and Blue. RGB are called primary colours because other colours can be obtained by linear combination of the three components red, green, and blue. Nevertheless, not all colours can be synthesized as combination of these three. A particular pixel can be represented by a three-dimensional vector, where (0, 0, 0) represents black, (k, k, k) is white, (k, 0, 0) is pure red, etc. Here, k is the quantization granularity for each component. This implies a colour space of $(k+1)^3$ distinct colours. For k = 255 there will be 2^{24} colours. The RGB model can be depicted as three-dimensional Cartesian co-ordinate system, where each point corresponds to a particular colour and intensity. Most cameras directly provide RGB values and thus make co-ordinate transformations unnecessary, which is of major advantage compared to some other colour spaces.

HSV. The HSV (Hue, Saturation, Value) colour space is a so-called perceptual colour space. (Note: HSV space is sometimes referred to as HSI for hue, saturation and intensity.) It represents colour in terms, which are more conform to colour understanding of human beings. The HSV model can be represented by a hexcone. Hue is given by the angle around the vertical axis relative to the 0° line, which is usually assigned to red. Saturation is denoted by the distance from the centre of the hexagon. On the outer edge, highly saturated colours are located, whereas at the centre, grey tones (saturation = 0) can be found. The intensity (value) is denoted by the vertical position in the hexcone. At the top of the hexcone, brightness is equal to 0, so that colour is black. Accordingly, the brightest colours are located at the hexcone's fat end. Usually, HSV is not provided by the camera itself, so transformation is necessary. The set of equations for the transformation can be found in [5] or in a different form in [6].

YIQ. For television systems, it was required to obtain the greyscale information only from one component. Therefore, for the United States television system standard NTSC (National Television System Committee) the YIQ colour space is used. It is composed of luminance (Y), hue (I), and saturation (Q). The first component Y provides all necessary information for a monochrome display. Moreover, it is possible to reduce the demand on bandwidth by limiting the bandwidth of the I and Q

signals without noticeable image degradation. However, the possibility of data reduction is not used in this work to ensure that the classifier has maximal information. The transformation from RGB to YIQ can be found e.g. in [7], but transformation is not necessary, if a NTSC camera is used. The geometrical model is a cuboid. Furthermore, YIQ exploits advantageous properties of the human visual system, especially the sensitivity to luminance.

YC_bC_r. The last colour space considered here, YC_bC_r, also contains a separate luminance component Y. Chrominance information is given by the difference signals B-Y = C_b and R-Y = C_r. This colour space is the basis of digital video and the European television standard PAL (Phase Alternating Line). The model of YC_bC_r can be represented by a cuboid, whose main diagonal is along the Y-axis. The advantage of this system for transmission also relies on the fact that the colour difference signals C_b and C_r can be subsampled to reduce bandwidth or data capacity. However, for the classification presented here no data reduction is used due to the before mentioned reason. The transformation from RGB to YC_bC_r can be found e.g. in [8], but also this kind of tristimulus values can be provided directly by a camera, which uses PAL standard.

Classification results. The classification results achieved by the neural network are rated by two criteria. First, the mean square error (MSE) was used as performance function. Additionally, we define

$$E_p = \sum_m \sum_n XOR(t_{m,n}, o_{m,n}), \tag{1}$$

where $t_{m,n}$ is a pixel of the reference (teacher) image and $o_{m,n}$ is the rounded pixel value of neural network's output image. m denotes the number of lines and n the number of columns respectively. E_p gives the number of misclassified pixels with respect to the reference image used during training. Fig. 1 shows the neural network's training and verification input image.

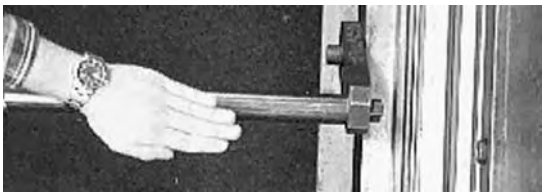

Fig. 1. Input image for training and verification

Single pixels have to be classified with respect to class ε_1 (output of the neural network is '1') or to class ε_2 (output of the neural network is '0'). Obviously, an increased complexity of the neural network will lead to better results for MSE and reduced number of misclassified pixels E_p (Table 1).

Furthermore, the results illustrate that the HSV colour space provides the best results with respect to MSE and E_p for all neural networks considered, except for the MSE of the 3/3/1 HSV network, where the MSE value is slightly higher than the corresponding values of the other colour spaces. However, the E_p value of the HSV network is the best for every network class considered. The different behaviour of the two quality measures can be explained by the fact that in the case of E_p binary output values occur whereas in the case of MSE float values are considered.

The graphical results of HSV for different neural networks are shown in the first row of Fig. 2. The hand and the parts of skin, respectively, are detected very well. However, some misclassifications are remaining, where pixels expected to be class ε_2 are classified as belonging to ε_1.

Nevertheless, the use of more complex network topologies does not provide result improvement proportional to the increase of complexity. For example, adding two additional hidden layers does not improve quality significantly, which can be seen by comparing image 2 and image 3 in the first row of Fig. 2. Still the misclassifications in the area of the background are the same. These results are valid for all colour spaces considered.

Table 1. **MSE and E_p (absolute and relative value, based on total number of pixels in training image) of three different multilayer feedforward networks for four different colour spaces**

NN Topology	Colour Space	MSE	E_p	E_p, %
3/3/1	RGB	0.0137031	484	1.80
	HSV	0.0137913	461	1.72
	YIQ	0.0137031	484	1.80
	YC_bC_r	0.0137116	495	1.84
3/3/4/1	RGB	0.0116042	430	1.60
	HSV	0.0115702	417	1.55
	YIQ	0.0117091	427	1.59
	YC_bC_r	0.0130913	449	1.67
3/3/4/8/1	RGB	0.0111774	404	1.50
	HSV	0.0102507	377	1.40
	YIQ	0.0115808	400	1.49
	YC_bC_r	0.0127204	431	1.60

In the second row of Fig. 2, the results for the most complex examined network structure are given which have been achieved for different colour spaces. It can be seen that there is nearly no difference between HSV, RGB, and YIQ, except YC_bC_r, whose result contains more misclassifications in the background. Obviously, complex network structures lead to better performance with respect to the number of misclassifications, nevertheless, post-processing is still required for eliminating artefacts in the background regions. Suppose there exists an extremely complex network structure that solves the problem accurately, however, computational effort would be much more higher compared to the combination of a simple network structure and low level post-processing.

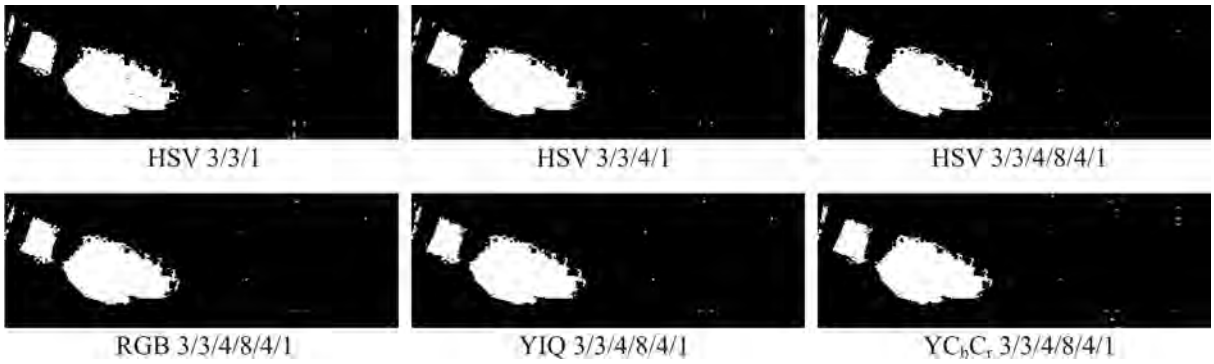

Fig. 2. Classification results of neural networks. In the first row, the results for three different neural networks in the case of HSV colour space are given. The results of the 3/3/4/8/4/1 network achieved for different colour spaces are given by the last image of the first row and the three images in the second row

Result enhancement

The classification results can be improved by binary morphological operations [9]. The computational effort associated with these algorithms is noticeable smaller than an additional layer with neurons. A binary image can be treated as a 2D point set, where points belonging to an object are pixels with value equal to one and points belonging to the background are pixels with value equal to zero. The two fundamental operations of morphology are dilation and erosion.

Binary dilation can be described as follows. If A represents all object pixels and SE represents a structure element, dilation is denoted by $A \oplus SE$ and is defined by

$$A \quad SE = \left\{ \mathbf{c} \in E^2 \,\middle|\, \mathbf{c} = \mathbf{a} + \mathbf{se}, \mathbf{a} \in A \text{ and } \mathbf{se} \in SE \right\}. \tag{2}$$

The dual operator of dilation is erosion, which is denoted by $A \ominus SE$ and is defined by

$$A \quad SE = \left\{ \mathbf{c} \in E^2 \,\middle|\, \mathbf{c} \ominus \mathbf{se} \in A \text{ for every } \mathbf{se} \in SE \right\}. \tag{3}$$

The duality of morphological operations is deduced from the existence of the set complement. Although dilation and erosion are dual operations, they are not inverse to each other. In particular, holes included in objects or indentations and projections cannot be reconstructed.

In our case, binary morphology is used to remove misclassifications in the neural network's output. At this step, the neighbourhood of pixels is considered as well. However, not the complete result image is needed to start processing but only the first n-1 lines of the result image and the first n pixel of line n, where n denotes the size of the n×n structure element. Thereby, low processing times can be achieved as the second processing unit can start without waiting until the first unit has finished processing of a whole frame. Output is a continuous data stream as well, from which a binary image can be reconstructed.

The example presented in Fig. 3 shows the successive use of binary morphological operations. Some misclassifications are included in the classification result of a 3/3/1 neural network (input presented in RGB space).These misclassifications can be eliminated by applying erosion and dilation with the same structure element (3×3). The differences to the reference image appear at the margins of the detected objects that are not completely covered. Covering of the hand's boundary areas can be obtained by using a larger dilation structure element. The precision of the shape will be slightly decreased, but the detected area will cover the hand completely. During simulation it has been observed that for this training image the processing time needed by a perceptron, which has an even more simple architecture (linear combiner followed by a hard limiter) than the neurons used here, is one third more than the time that is required by the combination of erosion and dilation applied to the neural network's binary output image.

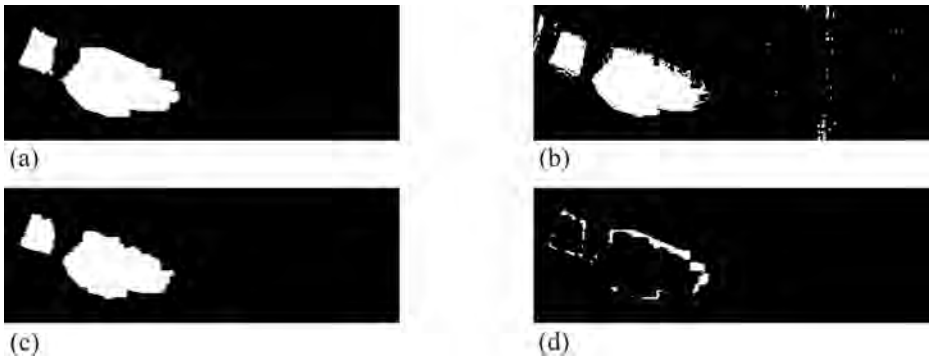

Fig. 3. Result enhancement by binary morphological operations. The reference image (a), the output of the neural network (b) (RGB 3/3/1), the result after morphological operations (c), and the difference between reference and enhanced result (d) are shown

However, to prove general results it has to be confirmed that the used neural network classifier is not specialised on the training pattern. As a matter of principle, learning strategies could lead to solutions, which only work well for inputs from the learning data set whereas other inputs could lead to misclassifications. Nevertheless, the neural network should be useable for many different images of the same acquisition situation and detection always should be possible. This is called generalisation. This aspect can be easily confirmed by presenting different images to the neural network and analysing the results.

In Fig. 4 (a) the binary output of the neural network (there are no morphological operations applied after classification) for a full frame (288×384 pixels) of the sequence from which the training pattern was taken is shown. It can be seen that the detection of both hands is possible.

In addition, the generalization behaviour has been tested with pairs of hands that have not been trained. The results achieved by the NN are also excellent (Fig. 4 (b), (c), (d)). Furthermore, the

classification errors in each binary output can be deleted by morphological operations, which have been discussed.

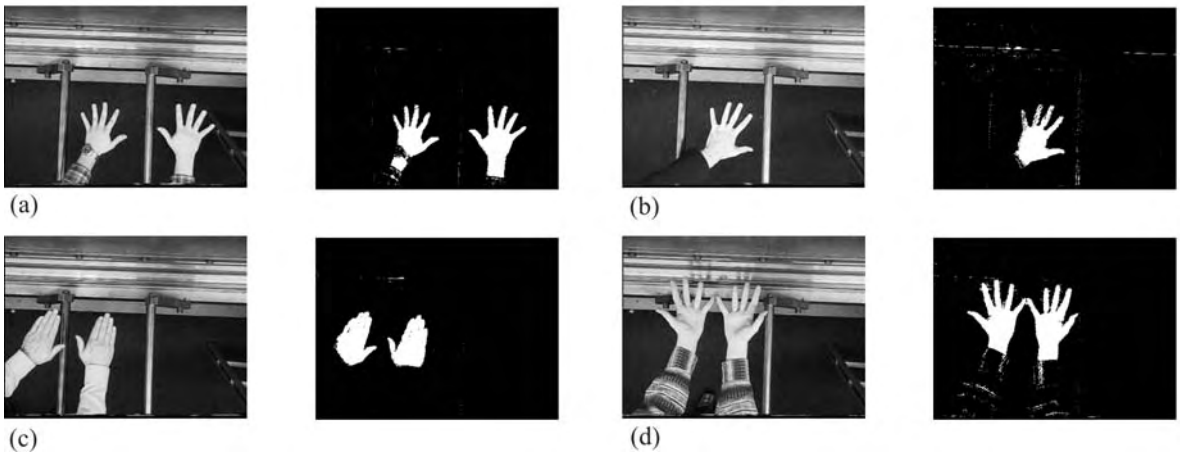

Fig. 4. Result of a full frame taken from the sequence from which the training pattern was obtained is shown in (a). Results of hand detection by the same neural network as in (a) for "non-trained hands" are illustrated in (b), (c), and (d). All results are achieved without any morphological enhancement

Conclusion

The combination of artificial neural networks and binary morphological operations leads to a fast colour detection system. The results reveal that a robust hand detection system can be obtained by this solution. Results presented in this work show that the choice of the colour space of the input source does not have major influence on the detection system. However, in order to obtain small processing times it should be recommended to use RGB, YIQ, or YC_bC_r, since they can be provided directly by digital cameras. In contrast to these, HSV colour space values only can be obtained via complex transformations from standard colour space representations. However, this will increase processing time. Furthermore, there is no need for complex neural network structure, since anyway; post processing has to be applied for removing misclassifications in the background regions. Therefore, a neural network that has the smallest possible size can be used for classification. Removing of remaining misclassifications can be done by erosion and dilation, which are so powerful that even the results of the 3/3/1 networks are sufficient. Although the precision of classified objects is not perfect, the system can provide a full coverage of the object (hand) by adding some kind of safe zone at margins without any additional processing step.

Best performance with respect to the processing time can be obtained by using dedicated hardware instead of using a general purpose computer for the proposed processing system. A hardware implementation of a suitable neural network with real time capability is presented in [2]. In addition, a fast hardware realization of morphological operations is possible as well [10].

References

1. *Haykin S.* Neural Networks - A Comprehensive Foundation. Macmillan College Publishing Company, New York, 1994.

2. *Krips M., Lammert T., Kummert A.* FPGA implementation of a neural network for a real-time hand tracking system // Proc. First IEEE International Workshop on Electronic Design, Test, and Applications.Christchurch, New Zealand, 2002. P. 313-317.

3. *Hagan M. T., Menhaj M. B.* Training Feedforward Networks with the Marquardt Algorithm // IEEE Trans. on Neural Networks. 1994. Vol.5, No.6. P. 989-993.

4. *Pratt W. K.* Digital Image Processing. 3rd ed. Wiley, New York, NY, 2001.

5. *Foley J. D., A. van Dam, Feiner S. K., Hughes J. F.* Computer Graphics – Principles and Practice. 2nd ed. Addison-Wesley, Reading, MA, 1990.

6. *Gonzalez R. C., Woods R. E.* Digital Image Processing. 2nd ed. Prentice Hall, Upper Saddle River, NJ, 2002.
7. *Sonka M., Hlavac V., Boyle R.* Image Processing, Analysis, and Machine Vision. 2nd ed. Brooks/Cole, Pacific Grove, CA, 1998.
8. *Poynton C. A.* A Technical Introduction to Digital Video. Wiley, New York, NY, 1996.
9. *Haralick R. M., Sternberg S. R., Zhuang X.* Image Analysis Using Mathematical Morphology // IEEE Trans. on Pattern Analysis and Machine Intelligence. 1987. Vol. PAMI-9, No.4. P. 532-550.
10. *Velten J. and Kummert A.* FPGA-based implementation of variable sized structuring elements for 2D binary morphological operations // Proc. First IEEE International Workshop on Electronic Design, Test, and Applications. Christchurch, New Zealand, 2002. P. 309-312.