FACE DETECTION USING LOCAL SMQT FEATURES AND SPLIT UP SNOW CLASSIFIER

Mikael Nilsson, Jörgen Nordberg, and Ingvar Claesson

Blekinge Institute of Technology
School of Engineering
Box 520, SE-372 25 Ronneby, Sweden
E-mail: mkn@bth.se, jno@bth.se, icl@bth.se

ABSTRACT

The purpose of this paper is threefold: firstly, the local Successive Mean Quantization Transform features are proposed for illumination and sensor insensitive operation in object recognition. Secondly, a split up Sparse Network of Winnows is presented to speed up the original classifier. Finally, the features and classifier are combined for the task of frontal face detection. Detection results are presented for the MIT+CMU and the BioID databases. With regard to this face detector, the Receiver Operation Characteristics curve for the BioID database yields the best published result. The result for the CMU+MIT database is comparable to state-of-the-art face detectors. A Matlab version of the face detection algorithm can be downloaded from http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=13701&objectType=FILE.

Index Terms— Object detection, Pattern recognition, Lighting, Image processing

1. INTRODUCTION

Illumination and sensor variation are major concerns in visual object detection. It is desirable to transform the raw illumination and sensor varying image so the information only contains the structures of the object. Some techniques previously proposed to reduce this variation are Histogram Equalization (HE), variants of Local Binary Patterns (LBP) [1] and the Modified Census Transform (MCT) [2]. HE is a computationally expensive operation in comparison to LBP and MCT, however LBP and MCT are typically restricted to extract only binary patterns in a local area. The Successive Mean Quantization Transform (SMQT) [3] can be viewed as a tunable tradeoff between the number of quantization levels in the result and the computational load. In this paper the SMQT is used to extract features from the local area of an image. Derivations of the sensor and illumination insensitive properties of the local SMQT features are presented.

Pattern recognition in the context of appearance based face detection can been approached in several ways [4, 5]. Techniques proposed for this task are for example the Neural Network (NN) [6], probabilistic modelling [7], cascade of boosted features (AdaBoost) [8], Sparse Network of Winnows (SNoW) [9], combination of AdaBoost and SNoW [2] and the Support Vector Machine (SVM) [10]. This paper proposes an extension to the SNoW classifier, the split up SNoW, for this classification task. The split up SNoW will utilize the result from the original SNoW classifier and create a cascade of classifiers to perform a more rapid detection. It will be shown that the number of splits and the number of weak classifiers can be arbitrary within the limits of the full classifier. Further, a stronger classifier will utilize all information gained from all weaker classifiers.

Face detection is a required first step in face recognition systems. It also has several applications in areas such as video coding, video conference, crowd surveillance and human-computer interfaces [5]. Here, a framework for face detection is proposed using the illumination insensitive features gained from the local SMQT features and the rapid detection achieved by the split up SNoW classifier. A description of the scanning process and the database collection is presented. The resulting face detection algorithm is also evaluated on two known databases, the CMU+MIT database [6] and the BioiD database [11].

2. LOCAL SMQT FEATURES

The SMQT uses an approach that performs an automatic structural breakdown of information. Our previous work with the SMQT can be found in [3]. These properties will be employed on local areas in an image to extract illumination insensitive features. Local areas can be defined in several ways. For example, a straight forward method is to divide the image into blocks of a predefined size. Another way could be to extract values by interpolate points on a circle with a radius from a fixed point [1]. Nevertheless, once the local area is defined it will be a set of pixel values. Let x be one pixel and $\mathcal{D}(x)$ be a set of $|\mathcal{D}(x)| = D$ pixels from a local area in an image. Consider the SMQT transformation of the local area

$$\mathrm{SMQT}_L : \mathcal{D}(x) \rightarrow \mathcal{M}(x) \tag{1}$$

which yields a new set of values. The resulting values are insensitive to gain and bias [3]. These properties are desirable with regard to the formation of the whole intensity image $\mathbf{I}(x)$ which is a product of the reflectance $\mathbf{R}(x)$ and the illuminance $\mathbf{E}(x)$ [12]. Additionally, the influence of the camera can be modelled as a gain factor g and a bias term b [2]. Thus, a model of the image can be described by

$$\mathbf{I}(x) = g\mathbf{E}(x)\mathbf{R}(x) + b. \tag{2}$$

In order to design a robust classifier for object detection the reflectance should be extracted since it contains the object structure. In general, the separation of the reflectance and the illuminance is an ill posed problem. A common approach to solving this problem involves assuming that $\mathbf{E}(x)$ is spatially smooth. Further, if the illuminance can be considered to be constant in the chosen local area then $\mathbf{E}(x)$ is given by

$$\mathbf{E}(x) = E, \forall x \in \mathcal{D}. \tag{3}$$

Given the validity of Eq. 3, the SMQT on the local area will yield illumination and camera-insensitive features. This implies that all

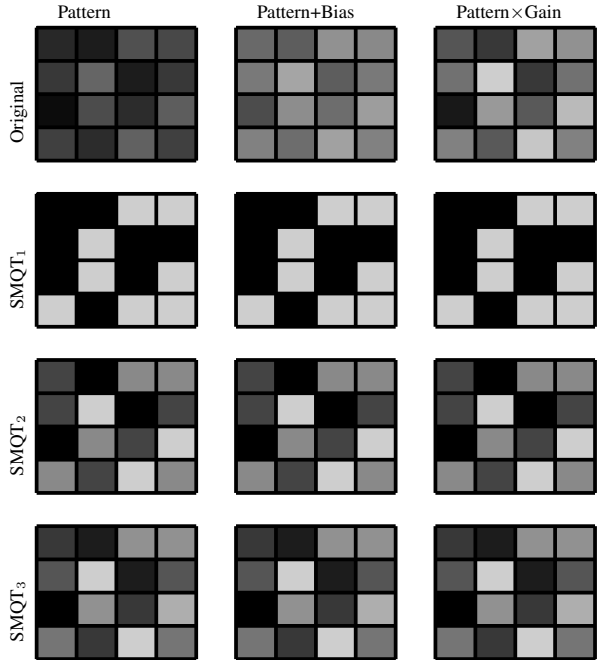

Fig. 1. Example 4×4 local area patterns and SMQT results.

local patterns which contain the same structure will yield the same SMQT features for a specified level L, see Fig. 1.

The number of possible patterns using local SMQT features will be $\left(2^L\right)^D$. For example the 4×4 pattern at $L = 1$ in Fig. 1 has $\left(2^1\right)^{4\times4} = 65536$ possible patterns.

3. SPLIT UP SNOW CLASSIFIER

The SNoW learning architecture is a sparse network of linear units over a feature space [9]. One of the strong properties of SNoW is the possibility to create lookup-tables for classification. Consider a patch $\mathcal{W}$ of the SMQT features $\mathcal{M}(\mathbf{x})$, then a classifier

$$\theta = \sum_{\mathbf{x}\in\mathcal{W}} h_x^{\text{nonface}}\left(\mathcal{M}(\mathbf{x})\right) - \sum_{\mathbf{x}\in\mathcal{W}} h_x^{\text{face}}\left(\mathcal{M}(\mathbf{x})\right) \tag{4}$$

can be achieved using the nonface table h_x^{nonface}, the face table h_x^{face} and defining a threshold for θ. Since both tables work on the same domain, this implies that one single lookup-table

$$h_x = h_x^{\text{nonface}} - h_x^{\text{face}} \tag{5}$$

can be created for single lookup-table classification.

Let the training database contain $i = 1, 2, \ldots, N$ feature patches with the SMQT features $\mathcal{M}_i(\mathbf{x})$ and the corresponding classes c_i (face or nonface). The nonface table and the face table can then be trained with the Winnow Update Rule [9]. Initially both tables contain zeros. If an index in the table is addressed for the first time during training, the value (weight) on that index is set to one. There are three training parameters; the threshold γ, the promotion parameter $\alpha > 1$ and the demotion parameter $0 < \beta < 1$. If $\sum_{\mathbf{x}\in\mathcal{W}} h_x^{\text{face}}\left(\mathcal{M}_i(\mathbf{x})\right) \leq \gamma$ and c_i is a face then promotion is conducted as follows

$$h_x^{\text{face}}\left(\mathcal{M}_i(\mathbf{x})\right) = \alpha h_x^{\text{face}}\left(\mathcal{M}_i(\mathbf{x})\right), \forall \mathbf{x} \in \mathcal{W}. \tag{6}$$

If c_i is a nonface and $\sum_{\mathbf{x}\in\mathcal{W}} h_x^{\text{face}}\left(\mathcal{M}_i(\mathbf{x})\right) > \gamma$ then demotion takes place

$$h_x^{\text{face}}\left(\mathcal{M}_i(\mathbf{x})\right) = \beta h_x^{\text{face}}\left(\mathcal{M}_i(\mathbf{x})\right), \forall \mathbf{x} \in \mathcal{W}. \tag{7}$$

This procedure is repeated until no changes occur. Training of the nonface table is performed in the same manner, and finally the single table is created according to Eq. (5).

One way to speed up the classification in object recognition is to create a cascade of classifiers [8]. Here the full SNoW classifier will be split up in sub classifiers to achieve this goal. Note that there will be no additional training of sub classifiers, instead the full classifier will be divided. Consider all possible feature combinations for one feature, $\mathcal{P}_i, i = 1, 2, \ldots, \left(2^L\right)^D$, then

$$v_x = \sum_{i=1}^{\left(2^L\right)^D} |h_x(\mathcal{P}_i)|, \quad \forall \mathbf{x} \in \mathcal{W} \tag{8}$$

results in a relevance value with respective significance to all features in the feature patch. Sorting all the feature relevance values in the patch will result in an importance list. Let $\mathcal{W}' \subseteq \mathcal{W}$ be a subset chosen to contain the features with the largest relevance values. Then

$$\theta' = \sum_{\mathbf{x}\in\mathcal{W}'} h_x\left(\mathcal{M}(\mathbf{x})\right) \tag{9}$$

can function as a weak classifier, rejecting no faces within the training database, but at the cost of an increased number of false detections. The desired threshold used on θ' is found from the face in the training database that results in the lowest classification value from Eq. (9).

Extending the number of sub classifiers can be achieved by selecting more subsets and performing the same operations as described for one sub classifier. Consider any division, according to the relevance values, of the full set $\mathcal{W}' \subseteq \mathcal{W}'' \subseteq \ldots \subseteq \mathcal{W}$. Then $\mathcal{W}'$ has fewer features and more false detections compared to $\mathcal{W}''$ and so forth in the same manner until the full classifier is reached. One of the advantages of this division is that $\mathcal{W}''$ will use the sum result from $\mathcal{W}'$. Hence, the maximum of summations and lookups in the table will be the number of features in the patch $\mathcal{W}$.

4. FACE DETECTION TRAINING AND CLASSIFICATION

In order to scan an image for faces, a patch of 32×32 pixels is applied. This patch is extracted and classified by jumping $\Delta x = 1$ and $\Delta y = 1$ pixels through the whole image. In order to find faces of various sizes, the image is repeatedly downscaled and resized with a scale factor $S_c = 1.2$.

To overcome the illumination and sensor problem, the proposed local SMQT features are extracted. Each pixel will get one feature vector by analyzing its vicinity. This feature vector can further be recalculated to an index

$$m = \sum_{i=1}^{D} \mathbf{V}(x_i)\left(2^L\right)^{i-1} \tag{10}$$

where $\mathbf{V}(x_i)$ is a value from the feature vector at position i. This feature index can be calculated for all pixels which results in the feature indices image.

A circular mask containing $P = 648$ pixels is applied to each patch to remove background pixels, avoid edge effects from possible filtering and to avoid undefined pixels at rotation operation, see Fig. 2.

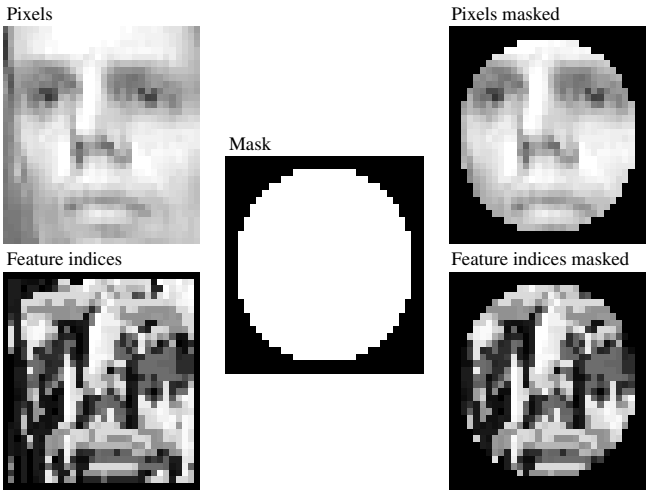

Fig. 2. Masking of pixel image and feature indices image. The features are here found by using a 3×3 local area and $L = 1$.

With the SNoW and the split up SNoW classifier, the lookup table is the major memory-intense issue. Consider the use of $N_{bit} = 32$ bit floating numbers in the table, then the classifier size (in bits) will be

$$S_{h_x} = N_{bit} \cdot P \cdot (2^L)^D. \tag{11}$$

Varying the size of the local area D and the level of the transform L directly affects the memory usage for the SNoW table classifier, see Tab. 1.

$D \downarrow \quad L \rightarrow$	1	2	3
2×2	40.5 KB	648 KB	-
3×3	1.26 MB	648 MB	324 GB
4×4	162 MB	10.1 TB	648 PB
5×5	81 GB	-	-

Table 1. Size of the classifier table with different local area sizes and different levels of the SMQT. $P = 648$ and $N_{bit} = 32$, see Eq. (11).

The choice of the local area and the level of the SMQT are of vital import to successful practical operation. For the split up SNoW classifier, with fast lookup table operation, one of the properties to consider is memory. Another is the local area required to make Eq. (3) valid. Finally, the level of the transform is important in order to control the information gained from each feature. In this paper, the 3×3 local area and level $L = 1$ are used and found to be a proper balance for the classifier. Some tests with 3×3 and $L = 2$ were also conducted. Although these tests showed promising results, the amount of memory required made them impractical, see Tab. 1.

The face and nonface tables are trained with the parameters $\alpha = 1.005$, $\beta = 0.995$ and $\gamma = 200$. The two trained tables are then combined into one table according to Eq. 5. Given the SNoW classifier table, the proposed split up SNoW classifier is created. The splits are here performed on 20, 50, 100, 200 and 648 summations. This setting will remove over 90% of the background patches in the initial stages from video frames recorded in an office environment.

Overlapped detections are pruned using geometrical location and classification scores. Each detection is tested against all other detections. If one of the area overlap ratios is over a fixed threshold, then the different detections are considered to belong to the same face. Given that two detections overlap each other, the detection with the highest classification score is kept and the other one is removed. This procedure is repeated until no more overlapping detections are found.

4.1. Face Database

Images are collected using a webcamera containing a face, and are hand-labelled with three points; the right eye, the left eye and the center point on outer edge of upper lip (mouth indication). Using these three points the face will be warped to the 32×32 patch using different destination points for variation, see Fig. 3. Currently, a grand total of approximatively one million face patches are used for training.

Fig. 3. Left - face image marked with three landmarks. Right - examples of how the three landmarks are used to warp the face to the 32×32 patches with different destination points for variation.

4.2. Nonface Database

Initially the nonface database contains randomly generated patches. A classifier is then trained using this nonface database and the face database. A collection of videos are prepared from clips of movies containing no faces and are used to bootstrap the database by analyzing all frames in the videos. Every false positive detection in any frame will be added to the nonface database. The nonface database is expanded using this bootstrap methodology. In final training, a total of approximatively one million nonface patches are used after bootstrapping.

5. RESULTS

The proposed face detector is evaluated on the CMU+MIT database [6] which contains 130 images with 507 frontal faces and the BioID database [11] which has 1521 images showing 1522 upright faces.

For the scanning procedure used here, the CMU+MIT database has 77138600 patches to analyze and the BioID database 389252799 patches. Both these databases are commonly used for upright face detection within the face detection community. The performance is presented with a Receiver Operation Characteristic (ROC) curve [13] for each database, see Fig. 4. With regard to the scanning used here, the False Positive Rate (FPR) is $1.93 * 10^{-7}$ and the True Positive Rate (TPR) is 0.95 if the operation on both databases is considered (77138600+389252799 patches analyzed).

The proposed local SMQT features and the split up SNoW classifier achieves the best presented BioID ROC curve and comparable results with other works on the CMU+MIT database. An extensive comparison to other works on these databases can be found in [2, 4, 5, 11].

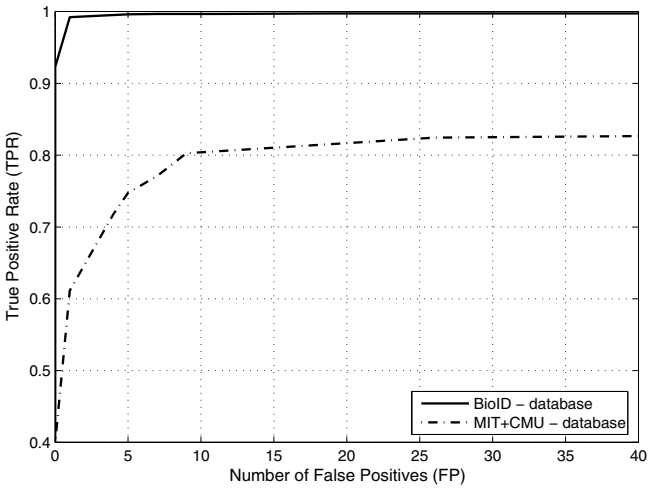

Fig. 4. Detection results on MIT+CMU (130 images, 507 faces to detect and 77138600 patches to analyze) and the BioID (1521 images, 1522 faces to detect and 389252799 patches to analyze) databases.

Note that the masking performed on each patch restricts detection of faces located on the edge of images, since important information, such as the eyes, can be masked away in those particular positions. This is typically the case with only few of the images found in the BioID database, hence to achieve a detection rate of one requires a large amount of false detections for those particular faces. The patches of size 32×32 also restrict detection of smaller faces unless upscaling is performed. The upscaling could be utilized on the CMU+MIT database, since it contains some faces that are of smaller size, however it is not considered here for the purpose of fair comparison with other works. Some of the faces were missed in the databases - a result which may have ensued due to scanning issues such as masking or patch size.

6. CONCLUSIONS

This paper has presented local SMQT features which can be used as feature extraction for object detection. Properties for these features were presented. The features were found to be able to cope with illumination and sensor variation in object detection.

Further, the split up SNoW was introduced to speed up the standard SNoW classifier. The split up SNoW classifier requires only training of one classifier network which can be arbitrarily divided into several weaker classifiers in cascade. Each weak classifier uses the result from previous weaker classifiers which makes it computationally efficient.

A face detection system using the local SMQT features and the split up SNoW classifier was proposed. The face detector achieves the best published ROC curve for the BioID database, and a ROC curve comparable with state-of-the-art published face detectors for the CMU+MIT database.

7. REFERENCES

[1] O. Lahdenoja, M. Laiho, and A. Paasio, "Reducing the feature vector length in local binary pattern based face recognition," in *IEEE International Conference on Image Processing (ICIP)*, September 2005, vol. 2, pp. 914–917.

[2] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, May 2004, pp. 91–96.

[3] M. Nilsson, M. Dahl, and I. Claesson, "The successive mean quantization transform," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005, vol. 4, pp. 429–432.

[4] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 1, pp. 34–58, 2002.

[5] E. Hjelmas and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 3, no. 3, pp. 236–274, 2001.

[6] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in *In Procedings of Computer Vision and Pattern Recognition*, June 1996, pp. 203–208.

[7] H. Schneiderman and T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '98)*, July 1998, pp. 45–51.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, vol. 1, pp. 511–518.

[9] D. Roth, M. Yang, and N. Ahuja, "A snow-based face detector," in *In Advances in Neural Information Processing Systems 12 (NIPS 12)*, pp. 855–861, MIT Press 2000.

[10] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines:an application to face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 1997, pp. 193–199.

[11] K. J. Kirchberg, O. Jesorsky, and R. W. Frischholz, "Robust face fetection using the hausdorff distance," in *Audio- and Video-Based Person Authentication - AVBPA 2001*, Josef Bigun and Fabrizio Smeraldi, Eds., Halmstad, Sweden, 2001, vol. 2091 of *Lecture Notes in Computer Science*, pp. 90–95, Springer.

[12] E. Adelson, "Lightness perception and lightness illusions," in *The Cognitive Neurosciences*, M. Gazzaniga, Ed. MIT Press, Cambridge, MA, 1999.

[13] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," Tech. Rep., HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto CA 94304, USA, 2004.