Weight-Based Facial Expression Recognition from Near-Infrared Video Sequences

Matti Taini, Guoying Zhao, and Matti Pietikäinen

Machine Vision Group, Infotech Oulu and Department of Electrical and Information Engineering,
P.O. Box 4500 FI-90014 University of Oulu, Finland
{mtaini,gyzhao,mkp}@ee.oulu.fi

Abstract. This paper presents a novel weight-based approach to recognize facial expressions from the near-infrared (NIR) video sequences. Facial expressions can be thought of as specific dynamic textures where local appearance and motion information need to be considered. The face image is divided into several regions from which local binary patterns from three orthogonal planes (LBP-TOP) features are extracted to be used as a facial feature descriptor. The use of LBP-TOP features enables us to set different weights for each of the three planes (appearance, horizontal motion and vertical motion) inside the block volume. The performance of the proposed method is tested in the novel NIR facial expression database. Assigning different weights to the planes according to their contribution improves the performance. NIR images are shown to deal with illumination variations comparing with visible light images.

Keywords: Local binary pattern, region based weights, illumination invariance, support vector machine.

1 Introduction

Facial expression is natural, immediate and one of the most powerful means for human beings to communicate their emotions and intentions, and to interact socially. The face can express emotion sooner than people verbalize or even realize their feelings. To really achieve effective human-computer interaction, the computer must be able to interact naturally with the user, in the same way as human-human interaction takes place. Therefore, there is a growing need to understand the emotions of the user. The most informative way for computers to perceive emotions is through facial expressions in video.

A novel facial representation for face recognition from static images based on local binary pattern (LBP) features divides the face image into several regions (blocks) from which the LBP features are extracted and concatenated into an enhanced feature vector [1]. This approach has been used successfully also for facial expression recognition [2], [3], [4]. LBP features from each block are extracted only from static images, meaning that temporal information is not taken into consideration. However, according to psychologists, analyzing a sequence of images leads to more accurate and robust recognition of facial expressions [5].

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 239–248, 2009.

Psycho-physical findings indicate that some facial features play more important roles in human face recognition than other features [6]. It is also observed that some local facial regions contain more discriminative information for facial expression classification than others [2], [3], [4]. These studies show that it is reasonable to assign higher weights for the most important facial regions to improve facial expression recognition performance. However, weights are set only based on the location information. Moreover, similar weights are used for all expressions, so there is no specificity for discriminating two different expressions.

In this paper, we use local binary pattern features extracted from three orthogonal planes (LBP-TOP), which can describe appearance and motion of a video sequence effectively. Face image is divided into overlapping blocks. Due to the LBP-TOP operator it is furthermore possible to divide each block into three planes, and set individual weights for each plane inside the block volume. To the best of our knowledge, this constitutes novel research on setting weights for the planes. In addition to the location information, the plane-based approach obtains also the feature type: appearance, horizontal motion or vertical motion, which makes the features more adaptive for dynamic facial expression recognition.

We learn weights separately for every expression pair. This means that the weighted features are more related to intra- and extra-class variations of two specific expressions. A support vector machine (SVM) classifier, which is exploited in this paper, separates two expressions at a time. The use of individual weights for each expression pair makes the SVM more effective for classification.

Visible light (VL) (380-750 nm) usually changes with locations, and can also vary with time, which can cause significant variations in image appearance and texture. Those facial expression recognition methods that have been developed so far perform well under controlled circumstances, but changes in illumination or light angle cause problems for the recognition systems [7]. To meet the requirements of real-world applications, facial expression recognition should be possible in varying illumination conditions and even in near darkness. Near-infrared (NIR) imaging (780-1100 nm) is robust to illumination variations, and it has been used successfully for illumination invariant face recognition [8]. Our earlier work shows that facial expression recognition accuracies in different illuminations are quite consistent in the NIR images, while results decrease much in the VL images [9]. Especially for illumination cross-validation, facial expression recognition from the NIR video sequences outperforms VL videos, which provides promising performance for real applications.

2 Illumination Invariant Facial Expression Descriptors

LBP-TOP features, which are appropriate for describing and recognizing dynamic textures, have been used successfully for facial expression recognition [10]. LBP-TOP features describe effectively appearance (XY plane), horizontal motion (XT plane) and vertical motion (YT plane) from the video sequence. For each pixel a binary code is formed by thresholding its neighborhood in a circle to the center pixel value. LBP code is computed for all pixels in XY, XT and YT planes or slices separately. LBP histograms are computed to all three planes or

slices in order to collect up the occurrences of different binary patterns. Finally those histograms are concatenated into one feature histogram [10].

For facial expressions, an LBP-TOP description computed over the whole video sequence encodes only the occurrences of the micro-patterns without any indication about their locations. To overcome this effect, a face image is divided into overlapping blocks. A block-based approach combines pixel-, region- and volume-level features in order to handle non-traditional dynamic textures in which image is not homogeneous and local information and its spatial locations need to be considered. LBP histograms for each block volume in three orthogonal planes are formed and concatenated into one feature histogram. This operation is demonstrated in Fig. 1. Finally all features extracted from each block volume are connected to represent the appearance and motion of the video sequence.

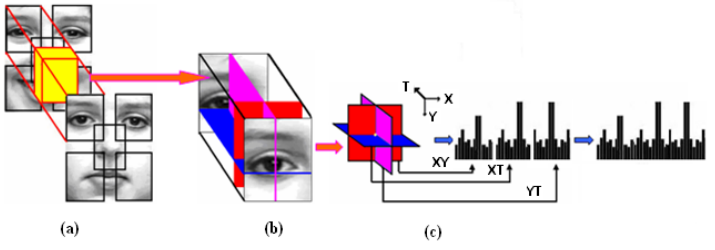

Fig. 1. Features in each block volume. (a) block volumes, (b) LBP features from three orthogonal planes, (c) concatenated features for one block volume.

For LBP-TOP, it is possible to change the radii in axes X, Y and T, which can be marked as R_X, R_Y and R_T. Also a different number of neighboring points can be used in the XY, XT and YT planes or slices, which can be marked as P_{XY}, P_{XT} and P_{YT}. Using these notations, LBP-TOP features can be denoted as $\text{LBP-TOP}_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$.

Uncontrolled environmental lighting is an important issue to be solved for reliable facial expression recognition. An NIR imaging is robust to illumination changes. Because of the changes in the lighting intensity, NIR images are subject to a monotonic transform. LBP-like operators are robust to monotonic gray-scale changes [10]. In this paper, the monotonic transform in the NIR images is compensated for by applying the LBP-TOP operator to the NIR images. This means that illumination invariant representation of facial expressions can be obtained by extracting LBP-TOP features from the NIR images.

3 Weight Assignment

Different regions of the face have different contribution for the facial expression recognition performance. Therefore it makes sense to assign different weights to different face regions when measuring the dissimilarity between expressions. In this section, methods for weight assignment are examined in order to improve facial expression recognition performance.

3.1 Block Weights

In this paper, a face image is divided into overlapping blocks and different weights are set for each block, based on its importance. In many cases, weights are designed empirically, based on the observation [2], [3], [4]. Here, the Fisher separation criterion is used to learn suitable weights from the training data [11].

For a C class problem, let the similarities of different samples of the same expression compose the intra-class similarity, and those of samples from different expressions compose the extra-class similarity. The mean ($m_{I,b}$) and the variance ($s^2_{I,b}$) of intra-class similarities for each block can be computed by as follows:

$$m_{I,b} = \frac{1}{C}\sum_{i=1}^{C}\frac{2}{N_i(N_i-1)}\sum_{k=2}^{N_i}\sum_{j=1}^{k-1}\chi^2\left(S_b^{(i,j)}, M_b^{(i,k)}\right) \quad , \tag{1}$$

$$s^2_{I,b} = \sum_{i=1}^{C}\sum_{k=2}^{N_i}\sum_{j=1}^{k-1}\left(\chi^2\left(S_b^{(i,j)}, M_b^{(i,k)}\right) - m_{I,b}\right)^2 \quad , \tag{2}$$

where $S_b^{(i,j)}$ denotes the histogram extracted from the j-th sample and $M_b^{(i,k)}$ denotes the histogram extracted from the k-th sample of the i-th class, Ni is the sample number of the i-th class in the training set, and the subsidiary index b means the b-th block. In the same way, the mean ($m_{E,b}$) and the variance ($s^2_{E,b}$) of the extra-class similarities for each block can be computed by as follows:

$$m_{E,b} = \frac{2}{C(C-1)}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\frac{1}{N_iN_j}\sum_{k=1}^{N_i}\sum_{l=1}^{N_j}\chi^2\left(S_b^{(i,k)}, M_b^{(j,l)}\right) \quad , \tag{3}$$

$$s^2_{E,b} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\sum_{k=1}^{N_i}\sum_{l=1}^{N_j}\left(\chi^2\left(S_b^{(i,k)}, M_b^{(j,l)}\right) - m_{E,b}\right)^2 \quad . \tag{4}$$

The Chi square statistic is used as dissimilarity measurement of two histograms

$$\chi^2(S,M) = \sum_{i}^{L}\frac{(S_i - M_i)^2}{S_i + M_i} \quad , \tag{5}$$

where S and M are two LBP-TOP histograms, and L is the number of bins in the histogram.

Finally, the weight for each block can be computed by

$$w_b = \frac{(m_{I,b} - m_{E,b})^2}{s^2_{I,b} + s^2_{E,b}} \quad . \tag{6}$$

The local histogram features are discriminative, if the means of intra and extra classes are far apart and the variances are small. In that case, a large weight will be assigned to the corresponding block. Otherwise the weight will be small.

3.2 Slice Weights

In the block-based approach, weights are set only to the location of the block. However, different kinds of features do not contribute equally in the same location. In LBP-TOP representation, the LBP code is extracted from three orthogonal planes, describing appearance in the XY plane and temporal motion in the XT and YT planes. The use of LBP-TOP features enables us to set different weights for each plane or slice inside the block volume. In addition to the location information, the slice-based approach obtains also the feature type: appearance, horizontal motion or vertical motion, which makes the features more suitable and adaptive for classification.

In the slice-based approach, the similarity within class and diversity between classes can be formed when every slice histogram from different samples is compared separately. $\chi^2_{i,j}(XY)$, $\chi^2_{i,j}(XT)$ and $\chi^2_{i,j}(YT)$ are the similarity of the LBP-TOP features in three slices from samples i and j. With this kind of approach, the dissimilarity for three kinds of slices can be obtained. In the slice-based approach, different weights can be set based on the importance of the appearance, horizontal motion and vertical motion features. Equation (5) can be used to compute weights also for each slice, when S and M are considered as two slice histograms.

3.3 Weights for Expression Pairs

In the weight computation above, the similarities of different samples of the same expression composed the intra-class similarity, and those of samples from different expressions composed the extra-class similarity. In that kind of approach, similar weights are used for all expressions and there is no specificity for discriminating two different expressions. To deal with this problem, expression pair learning is utilized. This means that the weights are learned separately for every expression pair, so extra-class similarity can be considered as a similarity between two different expressions.

Every expression pair has different and specific features which are of great importance when expression classification is performed on expression pairs [12]. Fig. 2 demonstrates that for different expression pairs, {E(I), E(J)} and {E(I), E(K)}, different appearance and temporal motion features are the most discriminative ones. The symbol "/" inside each block expresses the appearance, symbol "-" indicates horizontal motion and symbol "|" indicates vertical motion. As we can see from Fig. 2, for class pair {E(I), E(J)}, the appearance feature in block (1,3), the horizontal motion feature in block (3,1) and the appearance feature in block (4,4) are more discriminative and be assigned bigger weights, while for pair {E(I), E(K)}, the horizontal motion feature in block (1,3) and block (2,4), and the vertical motion feature in block (4,2) are more discriminative.

The aim in expression pair learning is to learn the most specific and discriminative features separately for each expression pair, and to set bigger weights for those features. Learned features are different depending on expression pairs, and they are in that way more related to intra- and extra-class variations of two specific expressions. The SVM classifier, which is exploited in this paper, separates

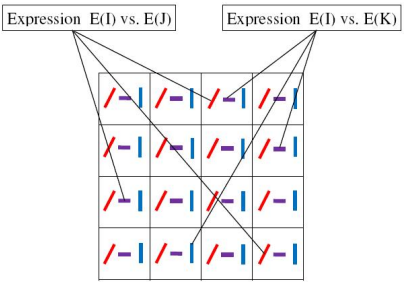

Fig. 2. Different features are selected for different class pairs

two expressions at a time. The use of individual weights for each expression pair can make the SVM more effective and adaptive for classification.

4 Weight Assignment Experiments

1602 video sequences from the novel NIR facial expression database [9] were used to recognize six typical expressions: anger, disgust, fear, happiness, sadness and surprise. Video sequences came from 50 subjects, with two to six expressions per subject. All of the expressions in the database were captured with both NIR camera and VL camera in three different illumination conditions: Strong, weak and dark. Strong illumination means that good normal lighting is used. Weak illumination means that only computer display is on and subject sits on the chair in front of the computer. Dark illumination means near darkness.

The positions of the eyes in the first frame were detected manually and these positions were used to determine the facial area for the whole sequence. 9×8 blocks, eight neighbouring points and radius three are used as the LBP-TOP parameters. SVM classifier separates two classes, so our six-expression classification problem is divided into 15 two-class problems, then a voting scheme is used to perform the recognition. If more than one class gets the highest number of votes, 1-NN template matching is applied to find out the best class [10].

In the experiments, the subjects are separated into ten groups of roughly equal size. After that a "leave one group out" cross-validation, which can also be called a "ten-fold cross-validation" test scheme, is used for evaluation. Testing is therefore performed with novel faces and it is subject-independent.

4.1 Learning Weights

Fig. 3 demonstrates the learning process of the weights for every expression pair. Fisher criterion is adopted to compute the weights from the training samples for each expression pair according to (6). This means that testing is subject-independent also when weights are used. Obtained weights were so small that they needed to be scaled from one to six. Otherwise the weights would have been meaningless.

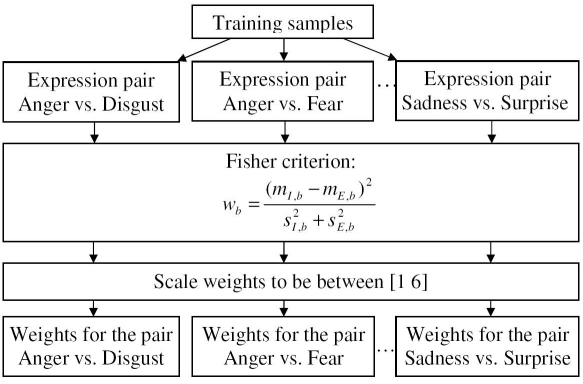

Fig. 3. Learning process of the weights

In Fig. 4, images are divided into 9 × 8 blocks, and expression pair specific block and slice weights are visualized for the pair fear and happiness. Weights are learned from the NIR images in strong illumination. Darker intensity means smaller weight and brighter intensity means larger weight. It can be seen from Fig. 4 (middle image in top row) that the highest block-weights for the pair fear and happiness are in the eyes and in the eyebrows. However, the most important appearance features (leftmost image in bottom row) are in the mouth region. This means that when block-weights are used, the appearance features are not weighted correctly. This emphasizes the importance of the slice-based approach, in which separate weights can be set for each slice based on its importance.

The ten most important features from each of the three slices for the expression pairs fear-happiness and sadness-surprise are illustrated in Fig. 5. The symbol "/" expresses appearance, symbol "-" indicates horizontal motion and symbol "|" indicates vertical motion features. The effectiveness of expression pair learning can be seen by comparing the locations of appearance features (symbol

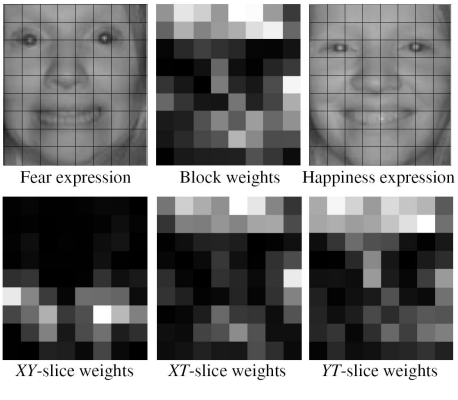

Fig. 4. Expression pair specific block and slice weights for the pair fear and happiness

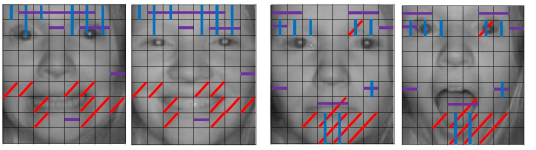

Fig. 5. The ten most important features from each slice for different expression pairs

"/") between different expression pairs in Fig. 5. For fear and happiness pair (leftmost pair) the most important appearance features appear in the corners of the mouth. In the case of sadness and surprise pair (rightmost pair) the most essential appearance features are located below the mouth.

4.2 Using Weights

Table 1 shows the recognition accuracies when different weights are assigned for each expression pair. The use of weighted blocks decreases the accuracy because weights are based only on the location information. However, different feature types are not equally important. When weighted slices are assigned to expression pairs, accuracies in the NIR images in all illumination conditions are improved, and the increase is over three percent in strong illumination. In the VL images, the recognition accuracies are decreased in strong and weak illuminations because illumination is not always consistent in those illuminations. In addition to facial features, there is also illumination information in the face area, and this makes the training of the strong and weak illumination weights harder.

Table 1. Results (%) when different weights are set for each expression pair

	Without weights	With weighted blocks	With weighted slices
NIR_Strong	79.40	77.15	82.77
NIR_Weak	73.03	76.03	75.28
NIR_Dark	76.03	74.16	76.40
VL_Strong	79.40	77.53	76.40
VL_Weak	74.53	69.66	71.16
VL_Dark	58.80	61.80	62.55

Dark illumination means near darkness, so there are nearly no changes in the illumination. The use of weights improves the results in dark illumination, so it was decided to use dark illumination weights also in strong and weak illuminations in the VL images. The recognition accuracy is improved from 71.16% to 74.16% when dark illumination slice-weights are used in weak illumination, and from 76.40% to 76.78% when those weights are used in strong illumination. Recognition accuracies of different expressions in Table 2 are obtained using weighted slices. In the VL images, dark illumination slice-weights are used also in the strong and weak illuminations.

Table 2. Recognition accuracies (%) of different expressions

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total
NIR_Strong	84.78	90.00	73.17	84.00	72.50	90.00	82.77
NIR_Weak	73.91	70.00	68.29	84.00	55.00	94.00	75.28
NIR_Dark	76.09	80.00	68.29	82.00	55.00	92.00	76.40
VL_Strong	76.09	80.00	68.29	84.00	67.50	82.00	76.78
VL_Weak	76.09	67.50	60.98	88.00	57.50	88.00	74.16
VL_Dark	67.39	55.00	43.90	72.00	47.50	82.00	62.55

Table 3 illustrates subject-independent illumination cross-validation results. Strong illumination images are used in training, and strong, weak or dark illumination images are used in testing. The results in Table 3 show that the use of weighted slices is beneficial in the NIR images, and that different illumination between training and testing videos does not affect much on overall recognition accuracies in the NIR images. Illumination cross-validation results in the VL images are poor because of significant illumination variations.

Table 3. Illumination cross-validation results (%)

Training	NIR_Strong	NIR_Strong	NIR_Strong	VL_Strong	VL_Strong	VL_Strong
Testing	NIR_Strong	NIR_Weak	NIR_Dark	VL_Strong	VL_Weak	VL_Dark
No weights	79.40	72.28	74.16	79.40	41.20	35.96
Slice weights	82.77	71.54	75.66	76.40	39.70	29.59

5 Conclusion

We have presented a novel weight-based method to recognize facial expressions from the NIR video sequences. Some local facial regions were known to contain more discriminative information for facial expression classification than others, so higher weights were assigned for the most important facial regions. The face image was divided into overlapping blocks. Due to the LBP-TOP operator, it was furthermore possible to divide each block into three slices, and set individual weights for each of the three slices inside the block volume. In the slice-based approach, different weights can be set not only for the location, as in the block-based approach, but also for the appearance, horizontal motion and vertical motion. To the best of our knowledge, this constitutes novel research on setting weights for the slices. Every expression pair has different and specific features which are of great importance when expression classification is performed on expression pairs, so we learned weights separately for every expression pair.

The performance of the proposed method was tested in the novel NIR facial expression database. Experiments show that slice-based approach performs better than the block-based approach, and that expression pair learning provides more specific information between two expressions. It was also shown that NIR

imaging can handle illumination changes. In the future, the database will be extended with 30 people using more different lighting directions in video capture. The advantages of NIR are likely to be even more obvious for videos taken under different lighting directions. Cross-imaging system recognition will be studied.

Acknowledgments. The financial support provided by the European Regional Development Fund, the Finnish Funding Agency for Technology and Innovation and the Academy of Finland is gratefully acknowledged.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Description with Local Binary Patterns: Application to Face Recognition. IEEE PAMI 28(12), 2037–2041 (2006)
2. Feng, X., Hadid, A., Pietikäinen, M.: A Coarse-to-Fine Classification Scheme for Facial Expression Recognition. In: Campilho, A.C., Kamel, M.S. (eds.) ICIAR 2004. LNCS, vol. 3212, pp. 668–675. Springer, Heidelberg (2004)
3. Shan, C., Gong, S., McOwan, P.W.: Robust Facial Expression Recognition Using Local Binary Patterns. In: 12th IEEE ICIP, pp. 370–373 (2005)
4. Liao, S., Fan, W., Chung, A.C.S., Yeung, D.-Y.: Facial Expression Recognition Using Advanced Local Binary Patterns, Tsallis Entropies and Global Appearance Features. In: 13rd IEEE ICIP, pp. 665–668 (2006)
5. Bassili, J.: Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face. Journal of Personality and Social Psychology 37, 2049–2059 (1979)
6. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Surveys 35(4), 399–458 (2003)
7. Adini, Y., Moses, Y., Ullman, S.: Face Recognition: The Problem of Compensating for Changes in Illumination Direction. IEEE PAMI 19(7), 721–732 (1997)
8. Li, S.Z., Chu, R., Liao, S., Zhang, L.: Illumination Invariant Face Recognition Using Near-Infrared Images. IEEE PAMI 29(4), 627–639 (2007)
9. Taini, M., Zhao, G., Li, S.Z., Pietikäinen, M.: Facial Expression Recognition from Near-Infrared Video Sequences. In: 19th ICPR (2008)
10. Zhao, G., Pietikäinen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE PAMI 29(6), 915–928 (2007)
11. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley & Sons, New York (2001)
12. Zhao, G., Pietikäinen, M.: Principal Appearance and Motion from Boosted Spatiotemporal Descriptors. In: 1st IEEE Workshop on CVPR4HB, pp. 1–8 (2008)