Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding xxx (2007) xxx–xxx
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Static topographic modeling for facial expression recognition and analysis

Jun Wang, Lijun Yin *

Department of Computer Science, State University of New York at Binghamton, Binghamton, NY 13902, USA

Received 6 September 2005; accepted 13 October 2006

Communicated by Mathias Kolsch

Abstract

Facial expression plays a key role in non-verbal face-to-face communication. It is a challenging task to develop an automatic facial expression reading and understanding system, especially, for recognizing the facial expression from a static image without any prior knowledge of the test subject. In this paper, we present a topographic modeling approach to recognize and analyze facial expression from single static images. The so-called topographic modeling is developed based on a novel facial expression descriptor, Topographic Context (TC), for representing and recognizing facial expressions. This proposed approach applies topographic analysis that treats the image as a 3D surface and labels each pixel by its terrain features. Topographic context captures the distribution of terrain labels in the expressive regions of a face. It characterizes the distinct facial expression while conserving abundant expression information and disregarding most individual characteristics. Experiments on person-dependent and person-independent facial expression recognition using two public databases (MMI and Cohn–Kanade database) show that TC is a good feature representation for recognizing basic prototypic expressions. Furthermore, we conduct the separability analysis of TC-based features by both a visualized dimensionality reduction example and a theoretical estimation using certain separability criterion. For an in-depth understanding of the recognition property of different expressions, the between-expression discriminability is also quantitatively evaluated using the separability criterion. Finally, we investigated the robustness of the extracted TC-based expression features in two aspects: the robustness to the distortion of detected face region and the robustness to different intensities of facial expressions. The experimental results show that our system achieved the best correct rate at 82.61% for the person-independent facial expression recognition.

Keywords: Facial expression representation; Facial expression recognition; Topographic context; Separability analysis

1. Introduction

Facial expression recognition and emotion analysis could help humanize computers and robots. Due to the wide range of applications in human–computer interaction, telecommunication, law enforcement and psychological research, facial expression analysis has become an active research area. Generally, an automatic facial expression reader consists of three major components: face detection, facial expression representation and facial expression classification [1,2]. Although the face detection component is a crucial part towards realizing an automatic expression recognition system, most research currently concentrates on how to achieve a better facial expression representation and feature extraction.

Facial expression representation concerns itself with the problem of facial feature extraction for modeling the expression variations with a certain accuracy and robustness. The Facial Action Coding System (FACS) [3,4], a psychological finding made over 25 years ago, is a typical example for representing and understanding human facial expressions. In order to read facial expressions correctly,

* Corresponding author.
E-mail address: lijun@cs.binghamton.edu (L. Yin).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.10.011

the action units have to be detected accurately and automatically. Based on the FACS, a number of systems were successfully developed for facial expression analysis and recognition [4,5,6]. Other systems have exploited the optical flow analysis to estimate the facial feature motions [7,8,9]. In addition, several other approaches are reported in recent literatures, such as Manifold-based analysis for blended expressions [10], Gabor-wavelet-labeled elastic graph matching for expression recognition [11], feature selection using the AdaBoost algorithm with classification by support vector machine [12], and learning Bayesian network for classification [13,14,15], etc. Impressive results were reported, however most representations of facial expressions are in a transformed domain with an implicit format. We believe that good features for representing facial expression could alleviate the complexity of the classification algorithm design. The facial expression descriptor should have the ability to decrease the significance of variations in age, gender and race when presenting the same prototypic expressions. In other words, it should work in a person-independent fashion.

Ideally, the facial expression can be modeled as a 3D face surface deformation actuated by the movement of the facial muscles. The intensity variations on a face image is caused by the face surface orientation and its reflectance. The resultant texture appearance provides an important visual cue to classify a variety of facial expressions [16]. Avoiding the ill-posed problem of 3D reconstruction from a single image, we exploit a so-called topographic representation to analyze facial expression on a terrain surface. Topographic analysis is based on the topographic primary sketch theory [17], in which the gray scale image is treated as a 3D terrain surface. Each pixel is assigned one type of topographic label based on the terrain surface structure. We can imagine that the facial skin ''wave'' is a reflection of a certain expression. Since the skin surface is represented by a topographic label ''map'', this ''map'' varies along with the change in facial expression. This fact suggests that topographic features can be expected to have the robustness associated with facial expression representation. It is thus of interest for us to investigate the relationship between these topographic features and the corresponding expressions in order to model the facial expression in an intuitive way.

Motivated by the topographic analysis technique [18,19], in this paper, we propose a novel facial expression descriptor—Topographic Context (TC)—to represent and classify facial expressions. Topographic Context describes the distribution of topographic labels in a region of interest of a face. We split a face image into a number of expressive regions. In order to obtain the topographic feature vector for an expression, the facial topographic surface is labeled to form a terrain map. Statistics on the terrain map is then conducted to derive the TC for each pre-defined expressive region. Finally, a topographic feature vector is created by concatenating all the TCs of expressive regions. With the extracted TC features, the facial expression can be recognized using some classification algorithms.

Note that most existing work have exploited time-varying features from video sequences for the dynamic facial expression analysis [5,8,9,12,15,20]. In this paper, we address the problem of facial expression representation and classification using static frontal-view facial images. This poses more of a challenge than using video sequences as no temporal information is available. Although temporal dynamics reflect facial behavior, it is essential to exploit the static image to represent and analyze configurational information of facial expressions [6]. The basic understanding of static facial expressions can help facilitate the application in psychological research and law enforcement when the temporal information is unavailable or inaccurate (e.g., due to the head motion). Our facial expression recognition system is tested on the static facial images from two facial expression databases, one is the commonly used Cohn–Kanade (CK) database [21] and the other is a newly published MMI database [22]. This system is completely person-independent, which means the subject to be tested has never appeared in the training set. In fact, person-independent expression recognition from a single static image is much more difficult due to the lack of prior information of the recognized subjects. The experiment shows that our TC-based expression representation has a good performance in classifying six universal expressions in terms of accuracy and robustness.

We conducted a detailed evaluation on the separability of the TC-based features, which includes an intuitive dimensionality reduction analysis and a theoretical analysis using a separability criterion. The evaluation results show that the TC-based expression features reflect intrinsic expression characteristics and decrease the variance on the age, race and subject. In order to evaluate the reliability of the TC-based feature representation, we further conducted the robustness analysis in terms of two aspects: robustness to the face region detection and robustness to the facial expression intensity. The experiments show that the TC-based feature representation has certain robustness to the distortion of facial landmark detection. Also, the TC-based features can be used to recognize mid-intensity facial expressions, although the performance is not as high as the extreme-intensity expression case.

The remainder of this paper is organized as follows. In Section 2, the topographic analysis is introduced. In Section 3, we present the concept of topographic context and give our static topographic facial expression model. The experiments on facial expression recognition are conducted in Section 4. The performance of the TC-based facial expression representation will be evaluated through the separability analysis in Section 5 and the robustness analysis in Section 6, followed by a discussion in Section 7. Finally, concluding remarks will be given in Section 8.

2. Topographic analysis

In order to derive the Topographic Context of facial images, we apply the topographic primal sketch theory

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

[17] to study the pixel characteristics in the face region. Topographic analysis treats the grey scale image as a terrain surface in a 3D space. The intensity $I(x,y)$ is represented by the height of the terrain at pixel (x,y). Fig. 1 shows an example of a face image and its terrain surface in the nose region. According to the property of the terrain surface, each pixel can be assigned one of the topographic labels: *peak*, *ridge*, *saddle*, *hill*, *flat*, *ravine*, or *pit* [17]. Hill-labeled pixels can be further specified as one of the labels *convex hill*, *concave hill*, *saddle hill* or *slope hill*. Saddle hills can be further distinguished as *concave saddle hill* or *convex saddle hill*. Saddle can be specified as *ridge saddle* or *ravine saddle*. So there are a total of 12 types of topographic labels [18,19]. In real face images the terrain surface of the face region is mainly composed of *ridge*, *ravine*, *convex hill*, *concave hill*, *convex saddle hill* and *concave saddle hill*. We illustrate these six topographic labels in Fig. 2, which will be used for our TC-based expression representation.

In order to calculate the topographic labels of the input gray scale image, a continuous surface $f(x,y)$ is used to fit the local $N \times N$ patch centered at (x,y) with the least square error. Then the first-order derivatives $\frac{\partial I(x,y)}{\partial x}$ and $\frac{\partial I(x,y)}{\partial y}$, and the second-order derivatives $\frac{\partial^2 I(x,y)}{\partial x^2}$, $\frac{\partial^2 I(x,y)}{\partial y^2}$ and $\frac{\partial^2 I(x,y)}{\partial x \partial y}$ are estimated using $f(x,y)$. Similar to the surface fitting approach used in [18], we use discrete Chebyshev polynomials up to the third degree as the bases spanning the vector space of these continuous functions. With the function $f(x,y)$, the partial derivatives can be approximated as

$$f^{(p,q)}(x,y) = \sum_{i=-N}^{N} \sum_{j=-N}^{N} I(x-i, y-j) h(i;p) h(j;q) \tag{1}$$

where $f^{(p,q)}(x,y)$ means the $(p+q)$th partial derivative at (x,y) with p along x axis and q along y axis. $h(i;p)$ and $h(i;q)$ are the smoothed differentiation filters from Chebyshev polynomials with degree p and q, respectively [23]. Thus the Hessian matrix is obtained as follows:

$$\mathbf{H}(x,y) = \begin{bmatrix} \frac{\partial^2 I(x,y)}{\partial x^2} & \frac{\partial^2 I(x,y)}{\partial x \partial y} \\ \frac{\partial^2 I(x,y)}{\partial x \partial y} & \frac{\partial^2 I(x,y)}{\partial y^2} \end{bmatrix} = \begin{bmatrix} f^{(2,0)}(x,y) & f^{(1,1)}(x,y) \\ f^{(1,1)}(x,y) & f^{(0,2)}(x,y) \end{bmatrix} \tag{2}$$

After applying eigenvalue decomposition to the Hessian matrix, we get:

$$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}} = [\mathbf{u}_1 \mathbf{u}_2] \cdot \mathrm{diag}(\lambda_1, \lambda_2) \cdot [\mathbf{u}_1 \mathbf{u}_2]^{\mathrm{T}} \tag{3}$$

where λ_1 and λ_2 are the eigenvalues and $\mathbf{u}_1$ and $\mathbf{u}_2$ are the orthogonal eigenvectors. The gradient magnitude $\|\nabla I(x,y)\|$ can be calculated as

$$\|\nabla I(x,y)\| = \sqrt{\left[\frac{\partial I(x,y)}{\partial x}\right]^2 + \left[\frac{\partial I(x,y)}{\partial y}\right]^2} \tag{4}$$

From the calculated λ_1, λ_2, $\mathbf{u}_1$, $\mathbf{u}_2$, $\|\nabla I(x,y)\|$ and the derivatives, the terrain labels can be assigned to pixel (x,y) by obeying a series of rules with some pre-defined empirical thresholds (e.g., T_g for gradients and T_λ for eigenvalues). For example, the ridge label is determined if the following condition is satisfied: $\|\nabla I(x,y)\| \leqslant T_g$, $\lambda_1 < -T_\lambda$ and $|\lambda_2| < T_\lambda$. The detailed labeling rules can be found in [19].

In order to reduce the influence of noise, smoothing preprocessing on the gray scale image is executed before calculating the derivatives. In our experiments, we use a Gaussian filter with a size of 15×15, a standard deviation $\sigma = 3.0$. An example of a face image and the smoothed 3D terrain surface of the nose region is shown in Fig. 1.

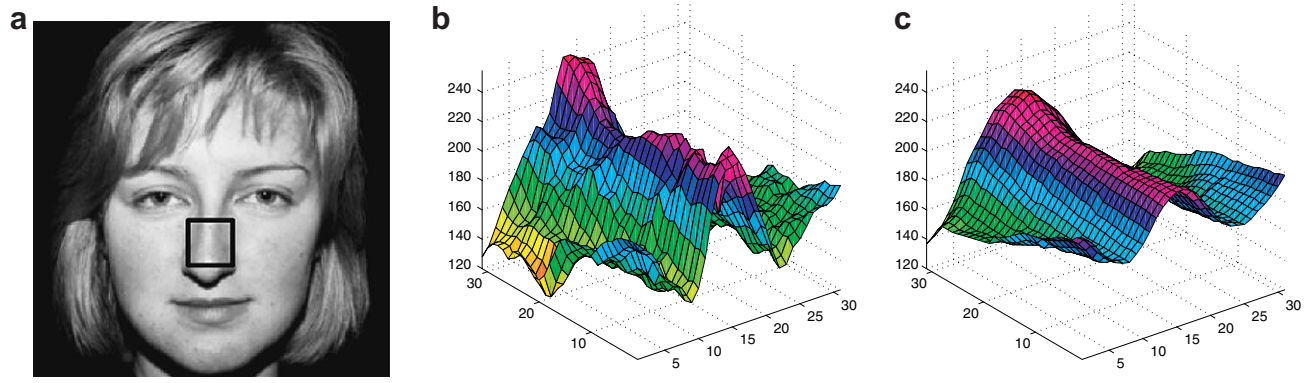

Fig. 1. Face image and the 3D terrain surface of the nose region. (a) Original face image; (b) terrain surface of the nose region of the original image; (c) terrain surface of nose region smoothed by a Gaussian filter.

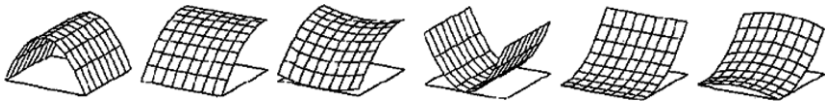

Fig. 2. A subset of the topographic labels [19]. The center pixel in each example carries the indicated label. From left to right, the labels are: *ridge*, *convex hill*, *convex saddle hill*, *ravine*, *concave hill* and *concave saddle hill*.

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

3. Topographic context based feature extraction

By applying topographic analysis, the original gray-scale image $I = \{I_{xy}\}$ is transformed to a so-called terrain map $T = \{T_{xy}\}$, which is composed of the terrain labels of each pixel.

Definition 3.1 (*Terrain Map*). A Terrain Map describes the topographic composition of a 3D terrain surface by indicating the type of terrain at each pixel.

In a terrain map, each pixel is represented by a certain topographic label. We survey the distribution of topographic labels on the terrain map by statistically analyzing 864 face images in the CK database. In real face images the terrain surface of the face region is mainly composed of *ridge*, *ravine*, *convex hill*, *concave hill*, *convex saddle hill* and *concave saddle hill*, which are more than 98.7% of all topographic label types. Table 1 shows our statistical results on 864 facial expression images. In the remaining part of the paper, only these six types of terrain labels are taken into account in our analysis.

The terrain map can be visualized using different colors to represent the different types of label. Fig. 3 shows the examples of facial expression images from the CK database and their corresponding terrain maps in the face regions. The terrain maps exhibit different patterns corresponding to different facial expressions. In order to give an explicit and quantitative description, we use a so-called Topographic Context to describe the statistical property of the terrain map.

Table 1
The ratios of 12 kinds of topographic labels in face image based on the statistic on 864 images from CK database

Convex hill	Convex saddle hill	Ridge	Flat	Ridge saddle	Peak
20.3%	31.8%	6.0%	0.0%	0.4%	0.4%
Concave hill	Concave saddle hill	Ravine	Slope	Ravine saddle	Pit
14.0%	21.8%	4.8%	0.0%	0.3%	0.2%

3.1. Topographic context (TC)

The motion of facial muscles due to changing expressions leads to a variation of facial terrain surface, which also results in the variation of image intensities. This fact suggests that topographic features can be expected to have the robustness for expression representation. To find an explicit representation for the fundamental structure of facial surface details, the statistical distributions of topographic labels within certain regions of interest are exploited.

Definition 3.2 (*Topographic Context*). Topographic context is a descriptor of the terrain features inside a region of interest of a gray-scale image. This descriptor is identified by the distribution of topographic labels in the corresponding terrain map.

In other words, the topographic context is a vector, whose elements are the ratios of the number of pixels with certain type of terrain label to the number of pixels in the whole region. For a given region in the terrain map, the topographic context can be computed as

$$\mathbf{e} = \left[\frac{n_1}{n}, \cdots, \frac{n_i}{n}, \cdots, \frac{n_t}{n}\right] \tag{5}$$

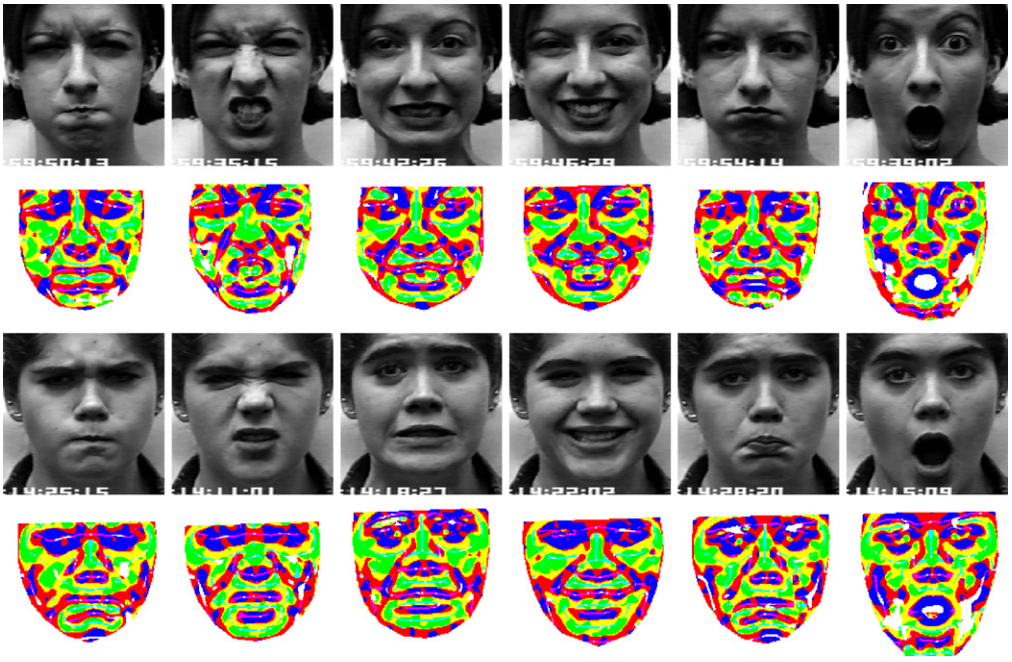

Fig. 3. Facial expression images and the corresponding terrain maps in face regions. From left to right, *Anger*, *Disgust*, *Fear*, *Happy*, *Sad* and *Surprise*.

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

where n_i is the number of pixels with the ith type of terrain label and $n = \sum_{i=1}^{t} n_i$ is the total number of pixels in the referred region. t denotes the number of terrain label types, which is 6 in our approach.

As an example shown in Fig. 4, we mark two square regions on the terrain map of a face image, and display the corresponding topographic contexts using normalized histograms. As illustrated in this figure, the topographic contexts reflect different characteristics of terrain features in different facial regions.

3.2. Face model

The formation of a facial expression is a combined actuation by a number of facial muscles. Based on the facial muscle distribution and the neuroanatomy of facial movement [24,25,26], we partition the face image into eight sub-regions, which are so-called *expressive regions*, so as to derive the topographic context efficiently.

Definition 3.3 (*Expressive Region*)**.** An expressive region is a pre-defined region of interest that reflects muscle actions and facial expressions. The expressive region may indicate facial organs or a single segment of the facial surface.

To locate the expressive regions in a face, we define a face model with 64 control points. We use the Active Appearance Model (AAM), which was well developed by Cootes etc. [27], to find the facial features and shapes. In the case of the high level of expression intensity, AAM may not detect the landmarks accurately. In order to evaluate our new facial expression features, the necessary manual adjustment is added to alleviate the influence of the facial landmark detection. In Section 6.1, we investigate the robustness of facial landmark detection with the artificial landmark distortions. As a result, the expressive regions can be constructed by polygons whose vertices are the detected facial landmarks, as shown in Fig. 5.

Note that not all of the facial areas are defined by the expressive regions. The nose bridge is excluded because of its lack of expressive information. Although the forehead surface may signify the expression-related furrows, due to the occasional occlusion by hairs, we exclude this region from our regions of interest. Currently, eight expressive regions are defined to construct the facial terrain model for TC-feature extraction.

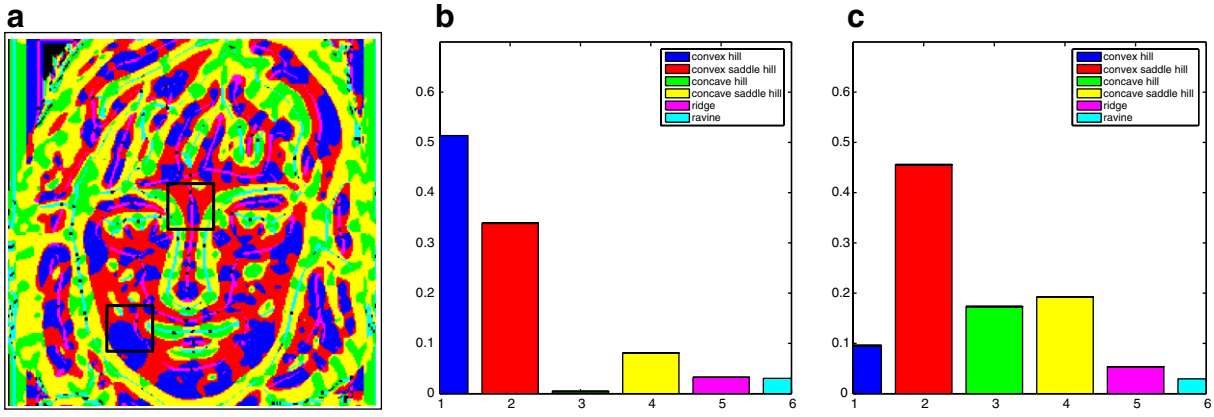

Fig. 4. Terrain map and topographic context by histogram representation. (a) Terrain map of the face image of Fig. 1; (b and c) topographic context of the jowl and glabella. The bars in the histogram (from left to right) represent *convex hill*, *convex saddle hill*, *concave hill*, *concave saddle hill*, *ridge* and *ravine*.

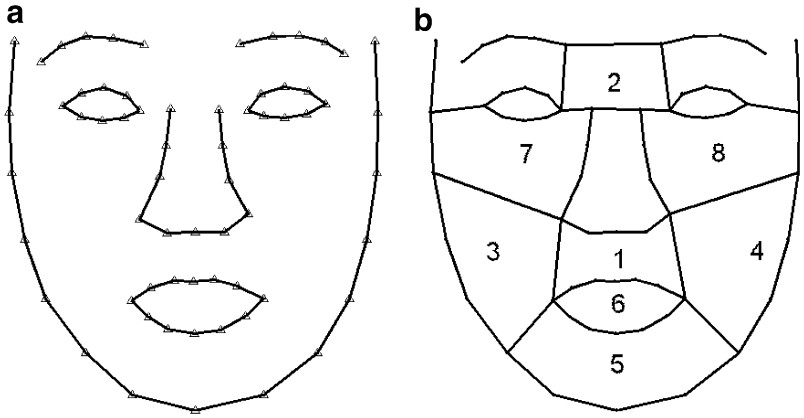

Fig. 5. Face model: (a) 64 facial landmarks; (b) 8 facial expressive regions.

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

3.3. TC-based facial expression feature

With the detected facial region or features, the expression feature can be derived. Here, we propose a novel feature, *Topographic Context* based expression feature, which has certain insensitivity to expression intensity and facial region locations.

The variation of facial expressions is a result of the motion of both facial organs and facial skin. For example, mouth opening is a distinct organ motion in some *surprise* expressions which present jowl surface stretching and flattening. Through the analysis of facial terrain surface, we are able to describe the terrain features in the expressive regions by topographic contexts. These topographic contexts will be eventually linked to a certain expression. The relationship between the topographic context and the expression is illustrated by an example in Fig. 6, which shows two subjects with three distinct prototypic facial expressions, *Surprise*, *Anger* and *Happy*. The corresponding TCs of expressive region 1 and region 6 are illustrated by their histograms.

Given the same facial expression from different subjects, the topographic contexts of expressive regions exhibit similar histogram characteristics. Fig. 6 illustrates the quantitative property of TCs for different subjects. It demonstrates that the topographic label distribution reflects different facial expressions. For example, in the expressive region 6, the *concave hill* and *ravine* rarely appear in mouth region when a face shows *anger* expression. Moreover, the prominent ratio of *convex saddle hill* usually results from happiness. Similarly, the prototypic expression information can also be uniquely represented in other seven regions.

Since the topographic context is defined in each expressive region, the eight expressive regions will produce eight topographic contexts. The combination of these TCs will generate a unique expression feature vector for a specific expression. In general, the procedure for generating expression feature vectors can be described as the following four steps:

1. Detect the 64 facial landmarks;
2. derive terrain map $T = \{T_{x,y}\}$ from original facial image $I = \{I_{x,y}\}$ by label each pixel based on its terrain property;
3. with the detected facial landmarks and the derived terrain map, compute topographic context for each expressive region;
4. combine the extracted topographic contexts of the eight expressive regions to construct the expression features.

As a result, the expression feature vector is constructed as

$$\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_k, \cdots, \mathbf{e}_M] \tag{6}$$

where M is the number of expressive regions ($M = 8$ in our experiment). The dimensionality of the expression feature vector is $t \times M$. Because only six kinds of terrain labels are used in our algorithm ($t = 6$), the expression feature of each face image is represented by a 48-dimensional vector, which is much lower than the dimensionality of the original image space.

4. Experiments on facial expression recognition

In this section, we conduct person-dependent and person-independent facial expression recognition experiments using the extracted TC-based features. Two different databases are used in our experiments. One is a newly created facial expression database (MMI-database) from the man–machine interaction group of Imperial College London [22] and the other is the widely used CK database [21]. MMI database mainly provide action units based facial expression data. Currently, only a few subjects with six universal expressions are available. The CK database consists of video sequences of subjects displaying distinct facial expressions, starting from neutral expression and ending with the peak of the expression. Because some subjects only show one or two facial expressions, we use a subset with 53 subjects for our experiment. On the average, four expressions appear for each subject. For each expression of a subject, the last four frames in the videos are selected. Notice that although we use several frames from the video sequence, we only treat these frames as static images for both training and testing without using any temporal information. Several samples of these two expression databases are shown in Figs. 3, 6 and 7. The summary of the two databases is presented in Table 1.

Although many existing classification algorithms could be employed for the experiment on facial expression recognition, our purpose is to evaluate the inherent discriminatory ability of the extracted TC features. The classifier design and training are not our emphasis in this paper. Hence, several standard and widely used classification approaches are used in our experiments (Table 2).

4.1. Person-dependent recognition test

In person-dependent test, first we divide the MMI and CK database into six and four subsets, respectively. Each subset contains all the subjects and each subject includes a set of available prototypic expressions. Then one of the subsets is selected for test while the remainder is used to construct the training set. It is a so-called "leave-one-out" cross-validation rule. The tests are repeated six times in the MMI database and four times in the CK database, with different test subset being used for each time. Three classifiers, quadratic discriminant classifier (QDC), linear discriminant analysis (LDA) and naive Bayesian network classifier (NBC), are employed to recognize the six prototypic facial expressions. Table 3 shows the average recognition rate. Table 4 reports the confusion matrix using the NBC classifier on the CK database.

As shown in Table 3, LDA achieves the highest recognition rate for the MMI database, and NBC does for the CK

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

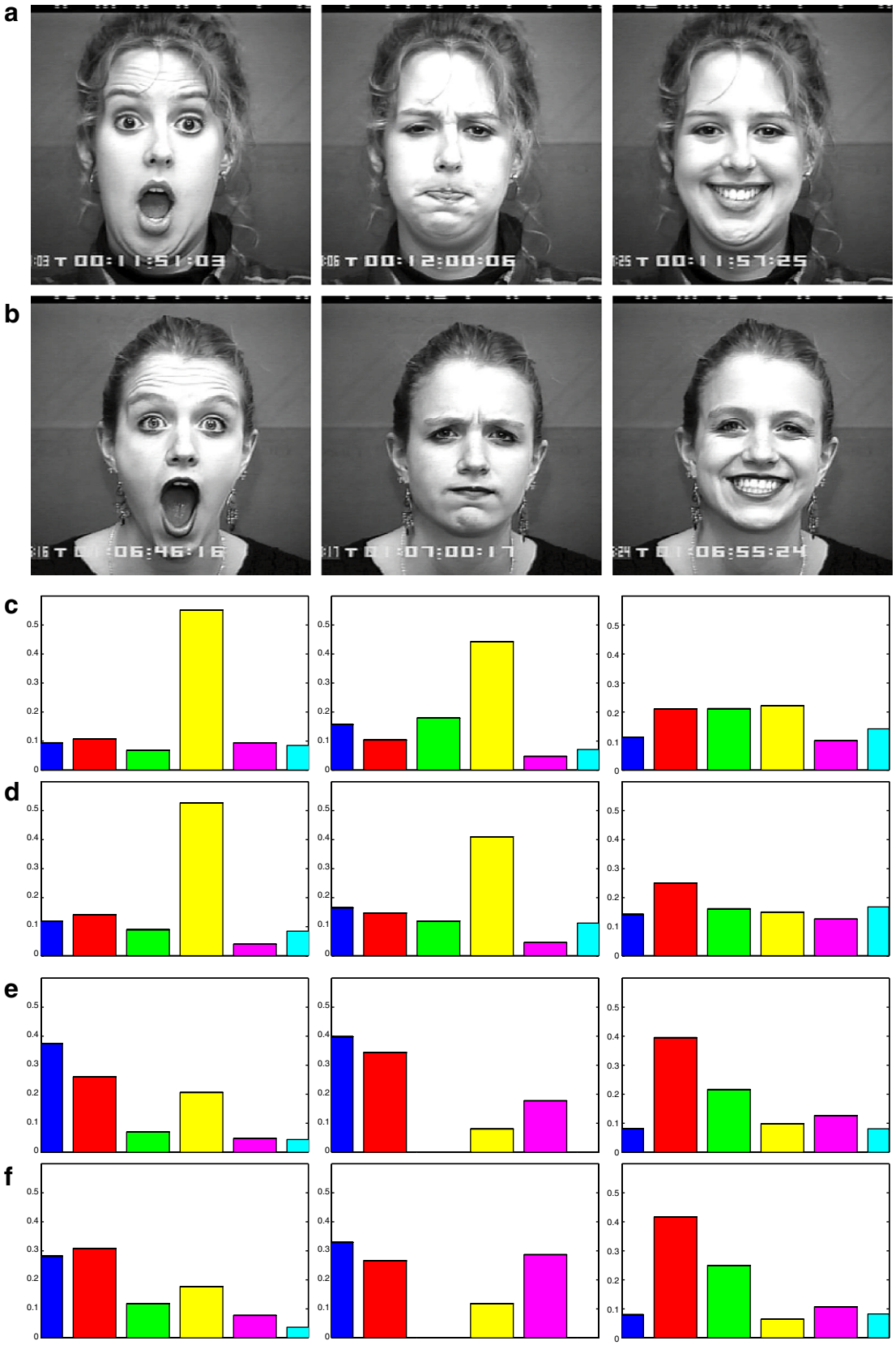

Fig. 6. TC features of two subjects with different expressions: (a and b) original images: *Surprise*, *Anger* and *Happy* of two subjects from the CK database; (c and d) corresponding TCs in the region 1 of subject (a) and (b); (e and f) corresponding TCs in the region 6 of subject (a) and (b).

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

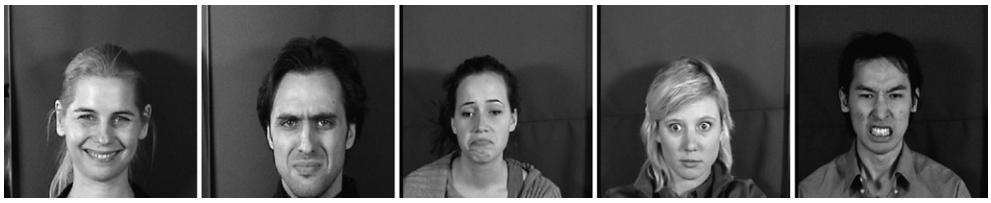

Fig. 7. Expression images from the MMI database.

Table 2
Summary of the databases used for our experiments

Database	# of subjects	# of images per subject for each expression	Overall # of images
MMI	5	6	180
CK	53	4	864

Table 3
Experimental results of person-dependent expression classification

Database\classifier	QDC (%)	LDA (%)	NBC (%)
MMI	92.78	93.33	85.56
CK	82.52	87.27	93.29

Table 4
Confusion matrix of the average case of NBC classifier on the CK database in person-dependent expression recognition

Input\output	Anger (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)
Anger	83.70	5.43	0.00	1.09	9.78	0.00
Disgust	2.78	93.06	0.00	0.00	2.08	2.08
Fear	1.92	3.85	83.65	7.69	2.89	0.00
Happy	0.00	0.50	1.00	98.50	0.00	0.00
Sad	1.47	0.74	0.00	0.00	97.06	0.74
Surprise	0.00	1.60	1.06	0.53	1.59	95.21

database. In Table 4, we can see that *Happy*, *Surprise*, *Disgust* and *Sad* are detected with high accuracy. *Fear* is sometimes confused with *Happy* while *Anger* is sometimes misclassified as *Sad*.

4.2. Person-independent recognition test

Note that the person-independent facial expression recognition is more challenging and necessary for practical applications. In order to train person-independent classifiers, we need more training subjects covering various patterns of same expression. Therefore, we use the CK database to carry out the person-independent tests. Because 53 subjects are contained in our database with a total of 864 images, we partition the whole set into 53 subsets, each of which corresponds to one subject and includes all the images of this subject with different expressions. The "leave-v-out" strategy is used for separating these subsets into training sets and test sets. This strategy is proved to be a more elaborate and expensive version of cross-validation [28]. In our experiment, the value of v is 10. It means that for each test, the database is partitioned into 43 subsets as the training set and 10 subsets as the test set. Note that any subject used for testing does not appear in the training set because the random partitioning is based on the subjects rather than the individual images. The tests are executed 20 times on each classifier with different partitions to achieve a stable recognition rate. The entire process guarantees every subject is tested at least once for every classifier. For each test, all the classifiers are reset and re-trained from the initial state. Totally four classifiers, including QDC, LDA, NBC and support vector classifier (SVC) with RBF kernel are used in the experiments.

Table 5
Experimental results of person-independent expression classification

Classifier	QDC	LDA	NBC	SVC
Recognition rate	81.96%	82.68%	76.12%	77.68%

Table 6
Confusion Matrix of the average case of LDA classifier for person-independent expression recognition

Input\output	Anger (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)
Anger	75.39	14.06	1.56	1.56	7.42	0.00
Disgust	12.50	69.50	10.04	0.00	5.87	2.08
Fear	3.72	2.93	68.88	21.54	1.33	1.60
Happy	0.00	2.70	4.46	91.76	1.08	0.00
Sad	13.20	5.28	1.94	0.18	79.40	0.00
Surprise	0.00	1.47	1.32	0.00	1.62	95.59

Table 5 reports the average recognition rates of these four classifiers, where LDA classifier achieve the highest accuracy with 82.68% correct rate. The confusion matrix of the average case for LDA classifier is shown in Table 6. The expressions *Surprise* and *Happy* are well detected with accuracy over 95% and 91%. *Anger*, *Disgust* and *Sad* are sometimes confused with each other, while *Fear* is sometimes misclassified as *Happy*.

5. Separability analysis

In order to further study the property of the TC-based expression representation, we investigate the inter-expression discriminability of TC-based expression features in this section. First of all, we examine the separability of the TC-based features in a low-dimensional space for an intuitive demonstration. Then, we give an in-depth theoretical analysis to quantitatively evaluate the between-expres-

sion separability using certain separability criterion, and show how the separability criterions used in this study agree with the results of human perception for distinguishing facial expressions.

5.1. Separability in a low-dimensional space: an intuitive example

High-dimensional TC-based expression features are hard to be visualized. It is intuitive for us to observe the clustering characteristic of the expression features in a low-dimensional space (i.e., 3D space). Here, we apply the Principal Component Analysis (PCA) to investigate the inter-expression discriminability of TC-based expression feature. To do so, we project the extracted TC-based features into a 3D space through the PCA dimensionality reduction.

Although a complex classification algorithm or a well-trained classifier can improve the performance, it heavily relies on the training data and lacks enough generalization ability. The unsupervised learning, for example, clustering characteristic gives a clear clue for an in-depth study on the separability of extracted features. We found that the TC-based facial feature exhibits a good inter-expression separability even in a low-dimensional space, especially for the distinct expressions *Happy* and *Surprise*. An example of separability study by reducing the dimensionality to a visualized space is given in Fig. 8. The TC features of 30 expression images (10 subjects, each of which has three expression images, *Happy*, *Surprise* and *Disgust*) are projected to a 3D space by PCA capturing 61.9% of the total energy. The selected subjects cover the variations of gender, race and face shapes.[1] As we can see, the output samples are roughly clustered in three sets which correspond to the three types of expressions. In general, the TC-based expression feature vectors exhibit a good ability in distinguishing expressions even in a very low-dimensional space.

For comparison, the original intensity image is also projected to the 3D space by applying PCA. As shown in Fig. 9, the three expression samples of each subject are distributed together. The three expressions are completely indistinguishable from projections of the original intensity images. The clustering sample intuitively demonstrates that the separability of TC-based expression feature is much superior to the intensity-based features, which exhibit the clustering property by subjects rather than by expressions.

Note that a visualized expression manifold by the Lipschitz embedding has been successfully used by Chang and Turk [10]. However, in [10], the constructed manifold is built for individual subject with a plenty of expression images of the *same* subject. Here, instead, we investigate the separability and clustering characteristic of extracted expression features using *multi*-subject and *multi*-expression data in a visualized 3D space. In general, the TC-based

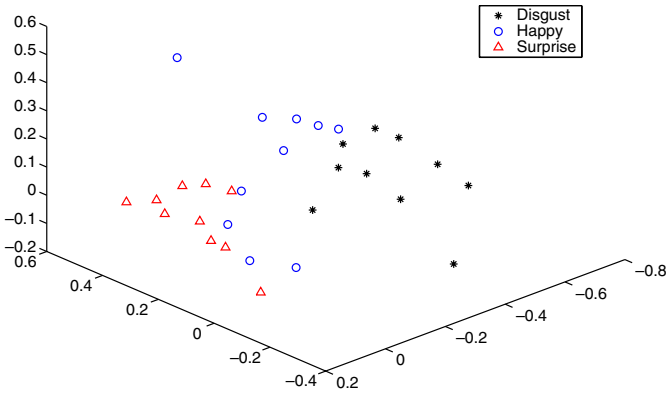

Fig. 8. 3D projections by applying PCA to reduce the dimensionality of TC-based expression features. The three expressions, *Happy*, *Surprise* and *Disgust* are roughly distinguishable.

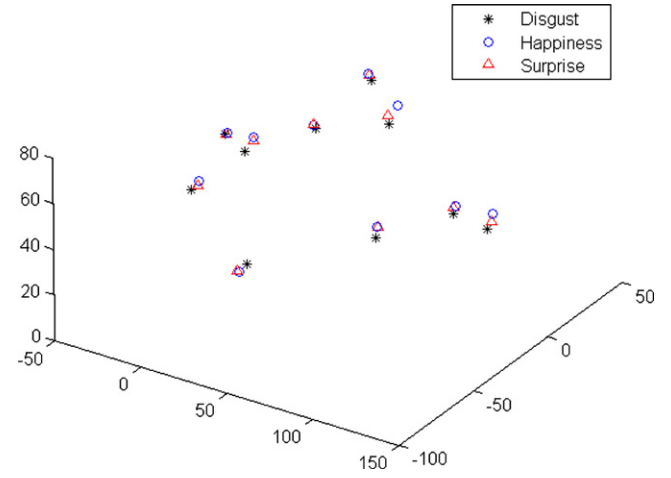

Fig. 9. 3D projections by applying PCA to reduce the dimensionality of intensity-based expression features. The three expressions, *Happy*, *Surprise* and *Disgust* are completely mixed together.

expression descriptor reflects more intrinsic expression property and alleviates the variations of non-expression factors, such as race and gender. Although the separability of other three expressions is hard to be visualized in 3D space, we can still distinguish them in a higher dimensional space, which is proved by the recognition experiments.

In the following section, we will give a theoretical analysis of the expression separability through the evaluation and comparison between the TC-based features and the intensity-based features.

5.2. Separability analysis of TC-based features

In order to quantitatively measure the separability of the expression features, a computable criterion should be defined first. There are some existing criteria, which are mainly used for feature selection [29,30], for example, probability based criterion, within-class and between-class distance based criterion, etc. Bayesian decision theory provides the probability based separability criterion (e.g., cor-

[1] The 10 selected subjects are numbered in the CK database as: S010, S011, S014, S022, S026, S050, S052, S055, S067 and S100.

rect probability or error probability), which can be used to evaluate and select the extracted features. In our multi-category case for expression classification, it is more efficient to compute the correct probability rather than the error probability. The correct probability metric is defined as

$$P(\text{correct}) = \sum_{i=1}^{c} P(x \in \mathscr{R}_i, \omega_i) = \sum_{i=1}^{c} \int_{\mathscr{R}_i} p(x|\omega_i) P(\omega_i)\, \mathrm{d}x \tag{7}$$

where $x \in \mathscr{R}_i$ means the feature space divided by the classifier, c is the number of classes, ω_i denotes the label of class and $P(\omega_i)$ is the prior probability of ω_i. Although the definition in Eq. (7) is simple, it is not feasible to calculate the correct probability in practice because it is difficult to estimate the *class-conditional probability density functions* $p(x|\omega_i)$ and the multiple integration has to be executed in high-dimensional space. Alternatively, the separability criterion based on within-class and between-class distance is more feasible and empirical. Suppose $\mathbf{x}_k^{(i)}$ and $\mathbf{x}_l^{(j)}$ are the d-dimensional features with the label ω_i and ω_j, respectively. The definition of average between-class distance in the case of multiple categories is as follow:

$$J_1(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \delta(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)}) \tag{8}$$

where, n_i, n_j are the numbers of samples in class ω_i and ω_j, P_i, P_j are the class-prior probabilities. $\delta(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)})$ denotes the distance between two samples, which is usually represented by the Euclidean distance

$$\delta(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)}) = (\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)})^{\mathrm{T}} (\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)}) \tag{9}$$

In order to represent the $J_1(x)$ in a compact form, two new concepts, *within-class scatter matrix* $\mathbf{S}_w$ and *between-class scatter matrix* $\mathbf{S}_b$, are introduced [29]

$$\mathbf{S}_w = \sum_{i=1}^{c} P_i \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{x}_k^{(i)} - \mathbf{m}_i)(\mathbf{x}_k^{(i)} - \mathbf{m}_i)^{\mathrm{T}} \tag{10}$$

$$\mathbf{S}_b = \sum_{i=1}^{c} P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^{\mathrm{T}} \tag{11}$$

where $\mathbf{m}_i$ is the mean of samples in the ith class

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{x}_k^{(i)} \tag{12}$$

$\mathbf{m}$ is the mean of all the samples.

$$\mathbf{m} = \sum_{i=1}^{c} P_i \mathbf{m}_i \tag{13}$$

With the definitions of Eqs. (10) and (11), we can get $J_1(x)$ in the following form [30,31]:

$$J_1(x) = \mathrm{tr}(\mathbf{S}_w + \mathbf{S}_b) \tag{14}$$

$J_1(x)$ is an efficient and computable separability criterion for feature selection. But it is not appropriate for comparing two different features because the value of calculated $J_1(x)$ depends on the scale and dimensionality of the feature space. Since the TC-based features and intensity-based features lie in two completely different spaces with different scales and dimensionalities, it is not easy to normalize the value of $J_1(x)$ for comparison. Here we use a comparable separability criterion to avoid the normalization. The new metric, $J_2(x)$, is defined as a natural logarithm of the ratio of determinant of within-class scatter matrix and between-class scatter matrix. $J_2(x)$ is intrinsically normalized in the comparable scale and reflects the separability of the features. For the value of $J_2(x)$, the larger the values are, the better the samples can be separated.

$$J_2(x) = \ln \frac{|(\mathbf{S}_b + \mathbf{S}_w)|}{|(\mathbf{S}_w)|} \tag{15}$$

We conducted the calculations of $J_2(x)$ on CK test set, which includes 864 facial expression images of 53 subjects. PCA is used to reduce the dimensionality of the extracted TC-based expression features. The selected separability metrics are calculated for each reduced dimensionality. For comparison, the similar experiment is also conducted using intensity features. Note that TC-based features and intensity features locate in completely different feature spaces, where the intensity feature has much higher dimensionality than the TC-based feature has. Therefore we normalize the calculated values by the percentage of retained energy rather than by the dimensionality. The facial expression separability and human subject separability are examined, respectively. Each examination is conducted by comparing the TC-based features and intensity features using the metric measurement $J_2(x)$. Fig. 10 shows the experimental results of the calculated value of $J_2(x)$. As shown in Fig. 10a, for the same ratio of retained energy, the value $J_2(x)$ of TC-based features is always higher than that of intensity-based features, which means that the TC-based features always exhibit much better between-expression separability than the intensity feature does.

The subject separability performance is also evaluated using the similar experimental strategy. Here, we re-label the facial images based on subjects rather than expressions. There are totally $c = 53$ classes in our test data. The estimated performance curve of separability criterion $J_2(x)$ as to the ratio of retained energy is shown in Fig. 10b. The separability curve of TC-based features is always lower than that of intensity-based features, which means that the intensity features have much better discriminability for subjects than the TC-based features have.

5.3. Between-expression separability analysis

The person-independent experiment in Section 4 (see Table 6) has shown that the correct recognition rates for different expressions are different. *Surprise* and *Happy* are the two most distinguishable expressions among the six prototype facial expressions, while *Anger*, *Disgust* and *Sad* are sometimes confused with each other. As a result, the confusion matrix of recognition exhib-

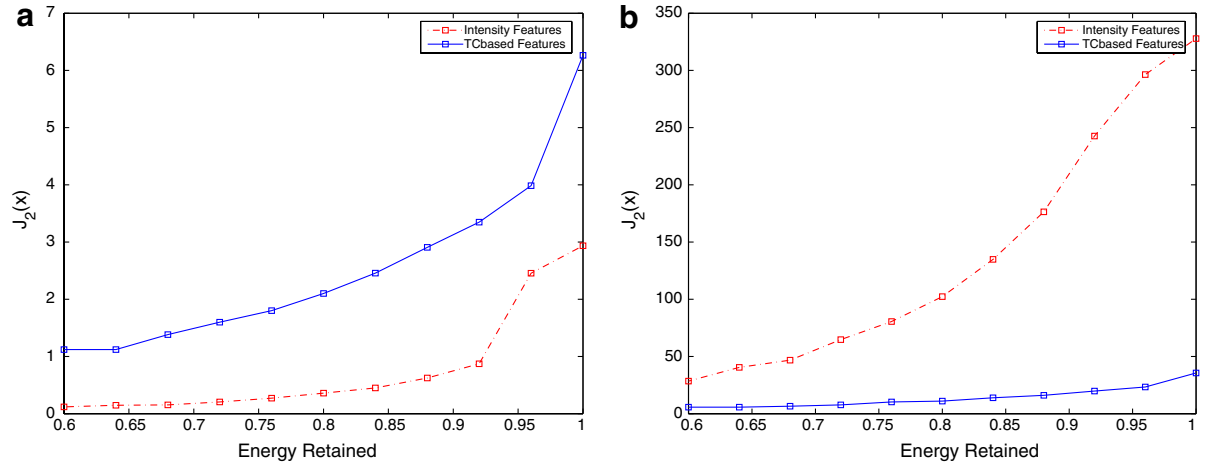

Fig. 10. Separability criterion ($J_2(x)$) comparison of TC-based features and intensity feature: (a) separability of expressions; (b) separability of subjects.

its a considerable imbalance. The automatic facial expression recognition system reported by Cohen etc. [20], which used different expression features and classifiers, shows the similar imbalance characteristic in classifying six prototype expressions. Here, we study the expression separability in order to provide insights into the intrinsic property of the facial expressions. We attempt to quantitatively explain the expression separability for better understanding the machine recognition and human cognition.

In order to evaluate the between-expression separability, the two-category separability criteria are used. Different from the criteria used in Section 5.2, here, the metrics for *within-class matrix* and *between-class scatter matrix* are specified for the two-category case. Assume that the six prototypic expressions are indexed from 1 to 6, we want to estimate the separability between expression e_i and e_j, where $1 \leqslant e_i,\ e_j \leqslant 6$ and $e_i \neq e_j$. In the test data set, we only use those images with these two types of expressions. The scatter matrices are calculated for the two-category case as

$$\mathbf{S}_w^{(e_i,e_j)} = \frac{1}{n}\left[\sum_{k=1}^{n_{e_i}}(\mathbf{x}_k^{(e_i)} - \mathbf{m}_{e_i})(\mathbf{x}_k^{(e_i)} - \mathbf{m}_{e_i})^{\mathrm{T}} + \sum_{l=1}^{n_{e_j}}(\mathbf{x}_l^{(e_j)} - \mathbf{m}_{e_j})(\mathbf{x}_l^{(e_j)} - \mathbf{m}_{e_j})^{\mathrm{T}}\right] \tag{16}$$

$$\mathbf{S}_b^{(e_i,e_j)} = P_{e_i}P_{e_j}(\mathbf{m}_{e_i} - \mathbf{m}_{e_j})(\mathbf{m}_{e_i} - \mathbf{m}_{e_j})^{\mathrm{T}} \tag{17}$$

where $n = n_{e_i} + n_{e_j}$ and $P_{e_i} + P_{e_j} = 1$. We estimate the separability criterion $J_2(x)$ for each pair of selected expressions while excluding other expressions. For each pair of selected expressions, the scatter matrices are computed using Eqs. (16) and (17). Then the separability criterion $J_2(x)$ is calculated. The evaluation is conducted on the extracted TC-based features with the entire energy conserved.[2]

[2] For a clear and distinguishable comparison, $e^{J_2(x)}$ is calculated instead of $J_2(x)$.

Table 7 illustrates the separability of all pairs of prototypic expressions, where the larger values of $J_2(x)$ indicates better discriminability between the two expressions. The results support the confusion matrix shown in Table 6. For example, expressions *Happy* and *Surprise* have the highest recognition rates. Conformably, these two expressions have a larger separability value $J_2(x)$ than other expressions. Also, expressions *Anger*, *Disgust* and *Sad* are easily misclassified because the separability value $J_2(x)$ is fairly small for any two of these three expressions.

In short, the recognition performance of the six prototypic expressions is mostly affected by the intrinsic appearance characteristics of expressions. The quantitative separability analysis is consistent with the recognition property for different expression, which is also obey the human vision perception rule.

6. Robustness analysis

6.1. Robustness to facial landmarks detection

As we know, the topographic context is defined on the individual expressive region of a face. It is conceivable that the performance of facial expression recognition is affected by the expressive region extraction, in other words, by the facial feature detection.

An accurate facial landmark localization algorithm could improve the performance of the expression recognition. Most previous work in facial expression analysis is under controlled conditions. For real applications, the possible complicated environment (e.g., complex background and uncontrolled illumination) could make the detection of facial landmarks inaccurate and unreliable.

In order to evaluate the TC-based feature as to how sensitive it is to the expressive region detection, we conduct two simulation experiments. In our first experiment, we simulate an uncontrolled environment by adding a Gaussian noise $k{\cdot}N(0,0.7)$ with a magnitude k to the detected landmarks. An example of the facial landmarks with differ-

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

Table 7
Confusion matrix of the expression separability criterion of $e^{J_2(x)}$

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	*N/A*	3.432	8.890	7.179	4.957	9.066
Disgust	3.432	*N/A*	5.357	8.290	4.192	8.544
Fear	8.890	5.357	*N/A*	4.570	8.400	7.225
Happy	7.179	8.290	4.570	*N/A*	12.490	15.395
Sad	4.957	4.192	8.400	12.490	*N/A*	9.335
Surprise	9.066	8.544	7.225	15.395	9.335	*N/A*

N/A, not applicable.

ent magnitudes of noise is shown in Fig. 11b and c. The second experiment simulates the noise through applying a random affine transformation to the detected entire facial shape. The transformation (i.e., rotation and translation) has the form as

$$\mathbf{M} = \mathbf{R}\left(\frac{k}{10} \cdot P(-0.5, 0.5)\right) \times \mathbf{T}(5k \cdot P(-0.5, 0.5)) \qquad (18)$$

where $P(-0.5, 0.5)$ is a uniform distribution between $[-0.5, 0.5]$ and k is the magnitude of noise. Fig. 11d–f shows the examples after affine transformation (Fig. 12).

For the training set, we use the clean facial landmarks. For each test image, the modified landmarks are used to calculate the facial expression features. The person-independent experiments described in Section 4.2 are repeated under the noise with different magnitudes. The performance curves of all four classifiers are recorded in Fig. 13. As shown in the figures, the recognition rate monotonically decreases as the noise magnitude increases, but it still maintains a fairly good performance, especially under the random Gaussian noise. The classifiers QDC and LDA maintain the correct recognition rate higher than 80% when $k = 4$. In short, the experimental results demonstrate that the TC-based facial expression features are not very sensitive to facial landmark detection. The main reason is that this feature representation is based on regional terrain label statistics. These statistical characteristics are not affected by local noises or by a certain shape distortion seriously.

Fig. 11. (a) Original landmarks; (b and c) landmarks added with random Gaussian noise with magnitude $k = 2$ and $k = 5$; (d–f) landmarks added after random affine transformation with magnitude $k = 1$, $k = 3$ and $k = 4$.

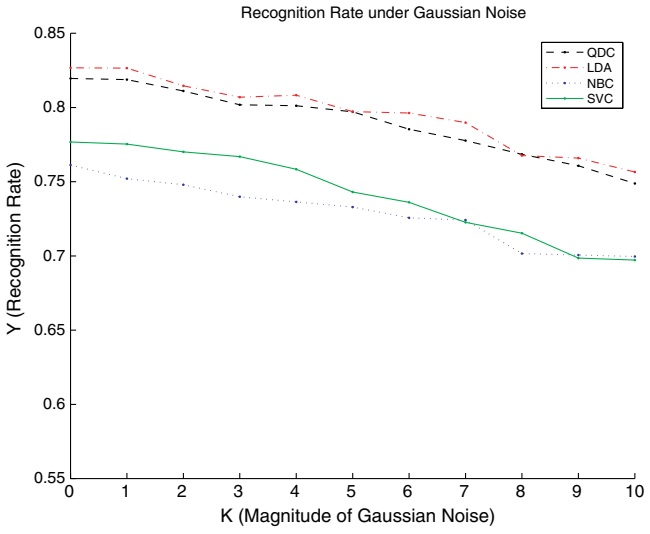

Fig. 12. Recognition under random Gaussian noise added to detected facial landmarks.

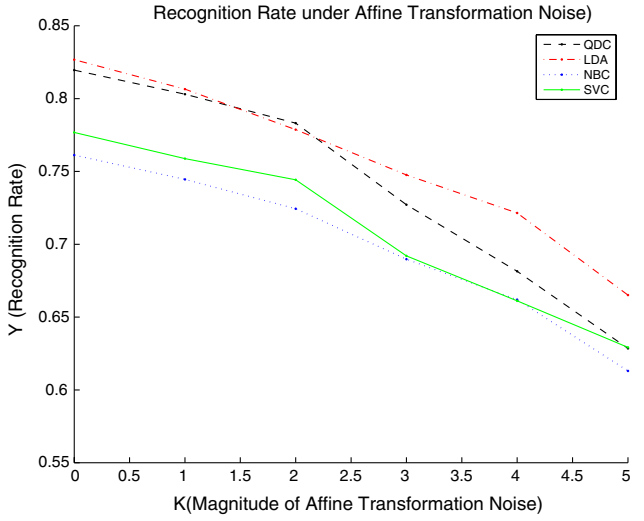

Fig. 13. Recognition under random affine transformation noise added to detected facial landmarks.

6.2. Robustness to expression intensity

Most of approaches for automatic facial expression analysis attempt to recognize a set of prototypic emotional expressions (e.g., fear, sadness, disgust, anger, surprise and happiness). In everyday life, however, such prototypic expressions with deliberate (i.e., posed) facial action rarely occur. Instead, the spontaneous expressions with subtle facial action appear more frequently [32]. Moreover, the posed facial action tasks typically differ in appearance and timing from the authentic facial expressions induced through events in the normal environment of the subject.

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

Therefore, investigation of recognition performance on different degrees of *expression intensity* (or degree of expression saturation) is of interest to us. Low-intensity expressions mostly reflect the spontaneity of human emotions. A database with authentic face expressions from 28 subjects was created by Sebe et al. [32], which includes four spontaneous expressions (neutral, happy, surprise and disgust). Bartlett et al. tested their algorithm on facial expression with different intensities measured by manual FACS coding [33]. Tian et al. combined Gabor features and neural network to discriminate eye status and compared their method with manual markers [34]. Although there is some existing work targeting the spontaneous expression analysis, so far, no standard definition and quantitative measurement of *expression intensity* has been reported.

As we know, recognizing a slight or low-intensity spontaneous expression is a very challenging task. Even for human vision system, it is difficult to interpret a slight or low-intensity expression based on only a single static image (e.g., because of the ambiguity). For the *static* image based universal expression recognition, most of previous research focused on facial images with extreme expressions or sufficiently perceivable expressions [1,35]. For low-intensity spontaneous expression recognition, it will be more effective to explore the temporal information through the analysis of dynamic expression sequences.

In the research on static image based facial expression recognition to date, little work has been done on the quantitative analysis of the recognition rate versus the expression intensity. In order to analyze the robustness and sensitivity of TC-based facial expression representation, we conduct an experiment to explicitly evaluate the performance of the extracted TC-based features under different degrees of expression intensities.

To do so, we extend the subset of CK database used in Section 4 by selecting more frames from the original video sequences. For each subject, we selected eight frames, which correspond to eight different degrees of perceived expressions. The selected frames are indexed from 1 to 8 based on the order of the frames in the video sequence. Because the expression video starts from the neutral expression to the extreme expression, the expression intensity increases from the index 1 to the index 8. As an example, eight facial expression images of one subject with different intensities are showed in Fig. 14. In our experiments, a total of 1672 face images, selected from 209 video sequences, are used.

The experimental strategy is same as the one used in Section 4.2. The recorded recognition performance is reported in Fig. 15. Moreover, the same criterion $J_2(x)$ that is used in Section 5 is applied to evaluate the between-expressions separability in different level of expression intensity. The calculated values of $J_2(x)$ corresponding to expression intensity indexed from 1 to 8 are recorded in Fig. 16.

Note that since there is no existing solution for the quantitative measurement of the expression intensity, our performance analysis in terms of expression intensity using frame indexes is approximate. Since different subjects may exhibit different styles of expressions (e.g., timing, speed, etc.), the frame-indexing based approach may not be able to extract the same expression intensity at a same indexed frame across different videos.

Experimental results on both expression recognition and between-expression separability illustrate the following properties of the TC-based facial expression recognition. First, the expression intensity does affect the recognition performance. Especially, when the expression intensity increases from index 1 to 5, the correct recognition rates are dramatically improved. In general, the recognition rate monotonically grows when the expression intensity increases. Second, when the expression reaches a certain degree of intensity, the correct recognition rate does not have a sig-

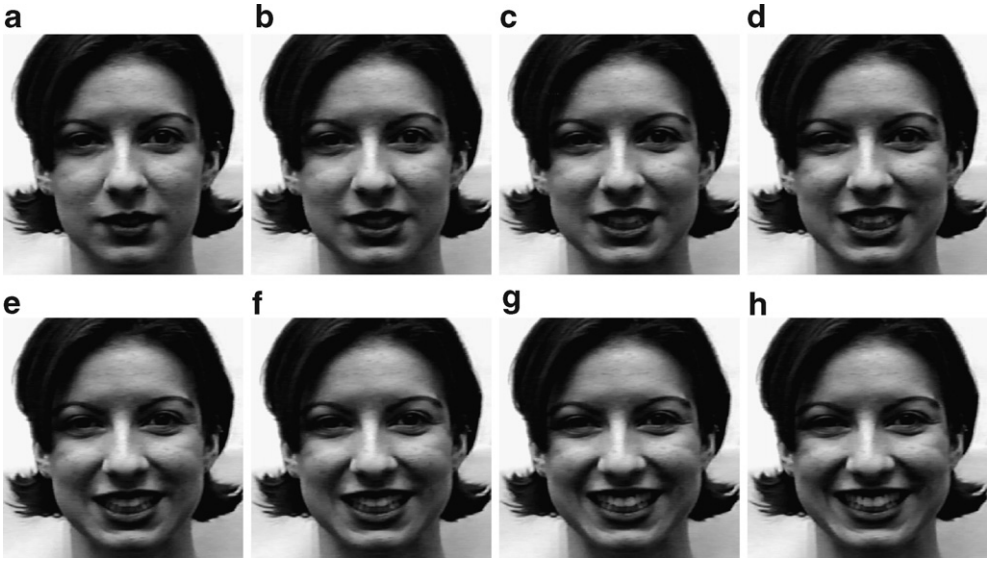

Fig. 14. Eight facial images with different expression intensities.

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

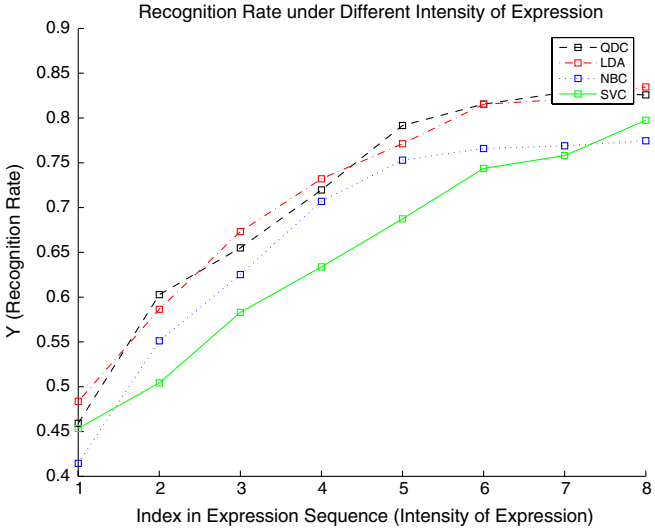

Fig. 15. Facial expression recognition performance under different facial expression intensity. X coordinate denotes the index of the test image in the selected video sequence, which displays the expression from low-intensity to extreme.

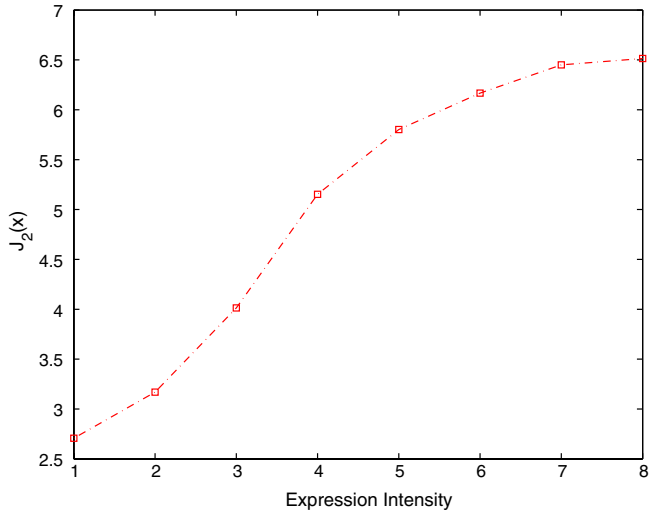

Fig. 16. Separability criterion ($J_2(x)$) under different expression intensities.

nificant increment. In other words, when expressions reach the "saturation" status with the high level of intensity, the performance of the recognition is not further improved.

7. Discussion

7.1. Advantages of TC-based feature representation

Expression feature extraction is a fundamental issue in automatic facial expression analysis and recognition. There are, in general, two categories of facial expression features used for classification, i.e., geometric features and appearance features [2]. Both feature representations are sensitive to imaging conditions and individual variations, such as face shape, scale and other factors. Although the normalization process could reduce the effect of imaging and individual variations, it is not trivial to obtain the normalized features without the human interaction and a neutral face input in some cases. For example, there is some prior work utilizing the line-based (or edge-based) caricatures to classify expressions [36]. Because the detected edge features are represented as binary levels (black and white), normalization is impossible. Moreover, because edges or corners are usually detected in the high contrast regions, it is hard to capture subtle facial features on the entire facial region. In [36], only three types of salient expressions can be classified correctly. However, in contrast, our topographic features are labeled with six different categories on each pixel of the entire face region. Since we use the statistics of the six types of topographic labels for each pre-defined facial region, the intrinsically normalized topographic context alleviates the influence of the individual variance, such as the face shape and scale. Although the calculation of TC is based on the detected facial landmarks, it is not sensitive to the facial region location because of its inherent statistical characteristics (proven by the experiments presented in Section 6.1). By contrast, geometric features are sensitive to the facial feature detection. Moreover, for the robustness analysis as to the head rotation, we compared the TC features with the Gabor-wavelet features on a synthesized facial expression image database using a recently published 3D facial expression database [37]. The experimental results reported in [38] show that the TC features are much more insensitive to facial pose than the Gabor-wavelet features.

It is worth noting that unlike most of existing approaches (e.g., FACS-based approaches), our TC-based facial expression representation does not require the high-resolution image input since the preprocessing of topographic labeling smooths the image details. Hence, it is conceivable that the TC-based features should have certain robustness to image quality.

7.2. Qualitative comparison with existing work

Identifying facial expression is a challenging problem in computer vision community. However, there is no baseline algorithm or standard database for measuring the merits of new methods, or determining what factors affect the performance. The unavailability of an accredited common database (e.g., like FERET database for face recognition) and evaluation methodology make it difficult to compare any new algorithm with the existing algorithms. As quoted by some researchers [1,6,35], most of existing work using static expressions (e.g., see Table 8 in [1] and Tables 2 and 3 in [35]) were tested on the in-house databases and some of the approaches are only evaluated by a person-dependent test. Although the CK database is widely used, some recent work [20,15] only utilized a portion of the database because not all the subjects in CK database have six prototypic expressions. This makes the algorithm comparison not fea-

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011

sible without knowing the exact test set used in their approaches.

Here, we give a brief qualitative comparison with the system recently reported by Cohen et al. in [20,15]. Both our work and the Cohen's system selected 53 subjects from the CK database for person-independent experiments. Although our system works on static images while Cohen's system worked on video sequences, both achieve similar recognition results and characteristics. Cohen's system achieved the best correct recognition rate at 81.80%, while our system achieves the best recognition rate at 82.68%. Both systems show the similar characteristics in classifying six prototypic expressions. For example, *Surprise* and *Happy*, as both reported, have the highest recognition rate. *Fear* is easily confused with *Happy* while *Sad* is confused with *Anger*.

8. Concluding remarks

In this paper we proposed a novel and explicit static topographic modeling for facial expression representation and recognition based on the so-called Topographic Context. Motivated by the mechanism of facial expression formation in the 3D space, we exploit the topographic analysis to study the facial terrain map which is derived by the topographic labeling techniques at the pixel level of detail. The facial expression is a behavior of an entire facial surface. The distribution of topographic labels described by the Topographic Context in expressive regions is a good reflection of this behavior. The performance of such a new facial expression descriptor is evaluated by the experiments of person-dependent and person-independent expression recognition on two public facial expression databases. The experimental results demonstrate that the TC-based expression features can capture the characteristics of facial expressions and achieve encouraging results in recognizing six prototypic expressions with person-independence. For a further investigation of the TC-based facial expression features, the separability analysis is conducted. The experimental results show that the TC-based features are appropriate for expression recognition because they reflect more intrinsic expression characteristics while alleviating the individual variations, such as race, gender and face shape. Robustness of TC-based features is evaluated in two aspects, the robustness to distortion of facial landmark detection and the robustness to different levels of expression intensities. Notice that the TC-based approach still has some limitations. Because the topographic feature estimation is based on the polynomial patch approximation, it is better than some edge detection approaches in terms of the illumination variations. However, this approach is still the pixel-based approach. It may not be robust to the dramatic illumination change. Due to the lack of the illumination-varied facial expression database, our future work is to develop a synthesis-based approach to simulate the various lighting conditions in order to study the illumination sensitivity.

As a challenging cross-disciplinary research topic, automatic facial expression recognition is still in its infancy and far from the real application. Although many advances and successes are reported in recent literature, many questions still remain open [2]. For the future work, we will work on a dynamic system using temporal information to improve the performance of the current static modeling. We will study the action units detection based on the topographic context domain. The integration of the location information and the topographic feature information could benefit the tracking of AUs. Moreover, the investigation on quantification and estimation of expression intensities is another future research direction.

Acknowledgments

The authors thank Dr. Jeffrey Cohn and the face group of CMU for providing the Cohn-Kanade database. We also thank Dr. Maja Pantic's group at Imperial College London for providing the MMI facial expression database. This work is supported in part by the National Science Foundation under Grants IIS-0414029 and IIS-0541044, and NYSTAR's James D. Watson Investigator program.

References

[1] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. PAMI 22 (2000) 1424–1445.
[2] S.Z. Li, A.K. Jain, Handbook of Face Recognition, Springer, New York, 2004.
[3] P. Ekman, W. Friesen (Eds.), The Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, San Francisco, 1978.
[4] G. Donato, M. Bartlett, J. Hager, P. Ekman, T. Sejnowski, Classifying facial actions, IEEE Trans. PAMI 21 (1999) 974–989.
[5] Y. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. PAMI 23 (2001) 1–9.
[6] M. Pantic, L. Rothkrantz, Facial action recognition for facial expression analysis from static face images, IEEE Trans. SMC-Part B: Cybern. 34 (2004) 1449–1461.
[7] K. Mase, Recognition of facial expression from optical flow, Proc. IEICE Trans. Spec. Issue Comput. Vis. Appl. 74 (1991) 3474–3483.
[8] Y. Yacoob, L. Davis, Recognizing human facial expression from long image sequences using optical flow, IEEE Trans. PAMI 16 (1996) 636–642.
[9] I. Essa, A. Pentland, Coding, analysis, interpretation, and recognition of facial expressions, IEEE Trans. PAMI 19 (1997) 757–763.
[10] Y. Chang, C. Hu, M. Turk, Probabilistic expression analysis on manifolds, in: CVPR, Washington DC, USA, 2004.
[11] M. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. PAMI 21 (1999) 1357–1362.
[12] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, in: CVPR Workshop on Face Processing in Video, Washington DC, USA, 2004.
[13] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, T. Huang, Learning bayesian network classifiers for facial expression recognition using both labeled and unlabeled data, in: CVPR, Wisconsin, USA, 2003.
[14] H. Gu, Q. Ji, Facial event classification with task oriented dynamic bayesian network, in: CVPR, Washington, DC, USA, 2004.
[15] I. Cohen, F.G. Cozman, N. Sebe, M.C. Cirelo, T.S. Huang, Semi-supervised learning of classifiers: theory, algorithms for bayesian

network classifiers and application to human–computer interaction, IEEE Trans. PAMI 26 (2004) 1553–1567.
[16] L. Yin, A. Basu, Generating realistic facial expressions with wrinkles for model based coding, Computer Vision and Image Understanding 84 (2001) 201–240.
[17] R. Haralick, L. Wasson, T. Laffey, The topographic primal sketch, Int. J. Robotics Res. 2 (1988) 91–118.
[18] L. Wang, T. Pavlidis, Direct gray-scale extraction of features for character recognition, IEEE Trans. PAMI 15 (1993) 1053–1067.
[19] O. Trier, T. Taxt, A. Jain, Ata capture from maps based on gray scale topographic analysis, in: The Third International Conference on Document Analysis and Recognition, Montreal, Canada, 1995.
[20] I. Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling, CVIU 91 (2003) 160–187.
[21] T. Kanade, J. Cohn,, Y. Tian, Comprehensive database for facial expression analysis, in: IEEE 4th International Conference on FGR, France, 2000.
[22] <http://www.mmifacedb.com/>.
[23] P. Meer, I. Weiss, Smoothed differentiation filters for images, J. Vis. Commun. Image Represent. 3 (1992) 58–72.
[24] W. Rinn, The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions, Psychol. Bull. 95 (1984) 52–77.
[25] L. Tassinary, J. Cacioppo, T. Geen, A psychometric study of surface electrode placements for facial electromyographic recording: 1. The brow and cheek muscle regions, Psychophysiology 26 (1989) 1–16.
[26] J. Cacioppo, D. Dorfman, Waveform moment analysis in psychophysiological research, Psychol. Bull. 102 (1987) 421–438.
[27] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. PAMI 23 (2001) 681–685.
[28] C. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
[29] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley-Interscience, New York, 2001.
[30] Z. Bian, X. Zhang, Pattern Recognition, Tsinghua University Press, Beijing, 1998.
[31] J.M. Lattin, J.D. Carroll, P.E. Green, Analysis Multivariate Data, Thomson Learning, New York, 2002.
[32] N. Sebe, M. Lew, I. Cohen, Y. Sun, T. Gevers, T. Huang, Authentic facial expression analysis, in: Proceedings of Automatic Face and Gesture Recognition, 2004.
[33] M. Bartlett, J. Hager, P. Ekman, T. Sejnowski, Measureing facial expressions by computer image analysis, Psychophysiology 36 (1999) 253–264.
[34] Y. Tian, T. Kanade, J. Cohen, Eye-state action unite detection by gabor wavelets, in: International Conference on Multimodal Interfaces, Beijing, China, 2000.
[35] B. Fasel, J. Luttin, Automatic facial expression analysis: survey, Pattern Recognit. 36 (2003) 259–275.
[36] Y. Gao, M. Leung, S. Hui, M. Tananda, Facial expression recognition from line-based caricatures, IEEE Trans. Syst. Man Cybern. A Syst. Hum. 33 (2003) 407–412.
[37] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, A 3D facial expression database for facial behavior research, in: Proceedings of Automatic Face and Gesture Recognition, Southampton, UK, April 2006.
[38] J. Wang, L. Yin, X. Wei, Y. Sun, 3D facial expression recognition based on primitive surface feature distribution, in: Proceedings of IEEE CVPR, New York, June 2006.

Please cite this article in press as: J. Wang, L. Yin, Static topographic modeling for facial expression recognition ..., Comput. Vis. Image Understand. (2007), doi:10.1016/j.cviu.2006.10.011