ABSTRACT

Facial expressions are often classified into one of several basic emotion categories. This categorical approach seems improper to treat faces with blended emotion, as well as hard to measure the intensity of an emotion. In this paper facial expressions are evaluated with dimensional approach of affect that was originally introduced by psycho-physiologic study. An expressional face can be represented as a point in a two-dimensional (2-D) emotional space characterized by arousal and valence factors. To link low-level face features with emotional factors, we propose a simple method that builds an "emotional mapping" by a coarse labeling on Cohn-Kanade database and a linear fitting on the labeled data. Our preliminary experimental result shows that the proposed emotional mapping can be used to visualize the distribution of affective content in a large face set and further retrieval expressional face images or relevant video shots by specifying a region in the 2-D emotional space.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Indexing methods; I.4.7 [Image Processing and Computer Vision]: Feature Measurement—Projections

General Terms
Algorithms, Design, Experimentation

Keywords
Facial expression analysis, Emotional space, Least square fitting, Image retrieval

1. INTRODUCTION

Nowadays more and more digital images and videos become available. With expanding data, there is also growing requirement of effective indexing and browsing the multimedia content. Facial expression can be one of important contents to index, since emotional expressions generally correspond to interesting events and highlights for video summarization and retrieval. Automatic facial expression analysis has been a hot research topic for years [11, 22]. This technique can be used to help user to search face images with specific expressions in large database, or retrieval exciting shots in long video.

Facial expressions are often evaluated by classifying face images into six basic emotion categories, i.e., happiness, sadness, anger, fear, surprise and disgust, which is based on a cross-culture study made by Ekman and Friesen in 1971 [9]. Limiting the evaluation of facial expression into predefined categories, however, may raise several problems. First, some expression can reflect multiple emotions, e.g., raised eyebrows and smiling mouth reveals a blend of surprise and happiness [22]. It is improper to classify such an expression as one of predefined basic emotions. Second, the intensity of an emotion is intrinsically ignored by the classification task itself. In this case, a small variation on face due to emotion may still be regarded as neutral face. Although, some recognition algorithm provides posterior probability of emotions [27, 8] that can be used as an intensity measurement. Third, facial expressions may not reflect any emotion at all, as they are simply facial actions. Considering this, some research [29, 30] analyzes facial expression by recognizing Action Units (AUs) [10], the visually observable facial muscle actions.

Besides categorical approach, dimensional approach has also been used in emotion evaluation, which can easily represent blended emotion and measure the intensity of emotion. One powerful emotional space is represented in terms of the three affective attributes, i.e., valence, arousal and control, which is known as 3-D VAC coordinate space [24]. Dietz and Lang [7] use this emotional space to develop an affective agent that serves as mediators between the computer and user. Yeasin et al. [32] use a similar 3-D space to analyze facial expression and measure levels of interest from video content. When Grimm et al. [13] study spontaneous facial expressions, they perform the evaluation of emotions both in basic emotion categories and in a 2-D emotional space (arousal vs. valence). Hanjalic and Xu [15] represent the affective video content as "affect curve" by linking emotional dimensions with low-level audio feature and the motion feature between consecutive video frames. Following them, Chan and Jones [2] explored a similar method for affective content retrieval in video. Note that in everyday interactions, people do not describe facial expressions as couple of real numbers. In the scenario of face image
3. EMOTIONAL MAPPING

face. The mapping is called whose coordinates characterize the emotion property of that professional face image can be represented as a point on this plane. So do “fear” and “angry”. This is determined by the affective factors.

However, our treatment will be convenient for illustration as well as for data processing. It is also non-negative value. However, the control dimension is often ignored in practical works [13, 15, 2] and the evaluation of emotion is limited in valence-arousal plane. Following them, we restrict our emotional mapping on this 2-D plane, too.

In order to facilitate the later labeling work on facial expression database, referred to [1], we indicate the rough area occupied by typical affect states on the valence-arousal plane, as shown in Figure 1.

2. THE 2-D EMOTIONAL SPACE

As mentioned in the Introduction, the emotion can be characterized by the intersection of three basic factors, i.e., valence, arousal and control. The terminology valence means the “type” of emotion, ranging continuously from pessimistic, neutral, to optimistic. Arousal stands for the “intensity” of emotion, ranging from sleepy, calm, to excited. Control is considered useful to distinguish between affective states with similar arousal and valence, ranging from no control to full control. However, it has been found that the effect of the control dimension appears quite small [7]. Also, it has also been shown that valence and arousal account for most of the independent variance in emotional responses [12]. For this reason, the control dimension is often ignored in practical works [13, 15, 2] and the evaluation of emotion is limited on valence-arousal plane. Following them, we restrict our emotional mapping on this 2-D plane, too.

In order to facilitate the later labeling work on facial expression database, referred to [1], we indicate the rough area occupied by typical affect states on the valence-arousal plane, as shown in Figure 1. Here both affective dimensions are bounded within a range of \([-1, +1]\) and the “neutral” state is placed at original point \((0, 0)\). This may a little conflict with the meaning of arousal that normally takes a non-negative value. However, our treatment will be convenient for illustration as well as for data processing. It is also noticed that “sad” and “disgust” lie adjacent on the Figure. So do “fear” and “angry”. This is determined by the affective factors.

The next step is to build a mapping so that an expression face image can be represented as a point on this plane whose coordinates characterize the emotion property of that face. The mapping is called emotional mapping.

3. EMOTIONAL MAPPING BUILDING

In order to build emotional mapping in a person-independent way, the training data should contain rich expressions posed by as many actors as possible. Among public available face datasets [23], we choose the Cohn-Kanade AU-coded facial expression database [17]. The released part of this database includes approximately 600 image sequences from 100 university students. Subjects were instructed by an experimenter to perform a series of facial displays that included single action units and their combinations. Each facial expression was captured in a sequence of images, starting from neutral face and ending at the maximum intensity of that expression. In the 2-D emotion space, the sequence should correspond to a trajectory departing from center and extending its way to surrounding region. Our aim is to derive this trajectory, or in other words, to estimate the emotion coordinates for all expression faces.

Since the Cohn-Kanade database is AU-coded, it is possible to know the emotion category of each face sequence using a finite set of rules of Emotion-AU correspondence [30, 31]. Though the rough region on valence-arousal plane can be estimated (Figure 1), the accurate coordinates are still hard to determine. Next, we show that a good estimation can be derived by a coarse classification on the image sequence and a linear fitting on the labeled data.

3.1 Coarse labeling on expression sequences

We make an assumption that the trajectory of an expression sequence exactly starts from original point and only extends along one of eight directions – positive/negative axes of coordinate system and the angle bisectors of quadrants. The last frame of sequence (apex frame) always reaches the boundary of unit square on the coordinate system. The points on trajectory are assumed to be distributed equally, whose coordinates are derived with linear interpolation. With these assumptions, the labeling process made easy, since the only thing to do for each sequence is picking up a proper direction out of eight choices. Further, the decision could be made by even more simple judgment at two affective dimensions: arousal is low, middle or high; valence is pessimistic, neutral or optimistic. Moreover, by utilizing the AU codes contained in the database, sequences with the same AU should follow the same trajectory on the emotional
3.2 Linear fitting on labeled data

Now we make a linear fitting on the coarsely labeled data. Let 2-D vector, $y$, represent the valence and arousal of a face image. Suppose $y$ has a linear relationship with face feature vector, $x$, as follows:

$$y = Ax + b$$

where $A$ is a transform matrix and $b$ is an offset vector. The formation of face feature $x$ will be specially explained in the next Section. With a set of $x$ and $y$, the linear relationship $A$ and $b$ can be derived with least square fitting. After the fitting, the coarsely labeled data in Figure 2 change their layout as shown in Figure 3. The face appearances according to the layout are shown in Figure 4. It can be seen that the faces are no longer concentrated on the axes or diagonals of coordinate system any more. They are now distributed over most area of the emotional plane, and more important, the faces are well laid out according to the meaning of valence and arousal!

3.3 Discussion on the emotional mapping

The promising result shown in Figure 3 and Figure 4 is simply derived by a linear fitting on coarsely labeled data. Thanks to the mapping function in Equation 1 whose continuity brings the faces with similar image feature come close to each other in the emotion space. The idea of data fitting between face feature and semantic factor was also adopted by Lanitis et al. [19] who modeled the aging variation and simulated aging effects on face images using linear, quadratic and cubic polynomials respectively. Intrinsically the aging is single factor variation whereas emotion is characterized by more factors.

In the Introduction of this paper, several limitations of categorical approach were mentioned. Now, it would be easier to understand that these limitations could be overcome or relieved by the dimensional approach in some extent. Mixture of emotions, intensity of emotion and even a subtle variation on face, all they are now able to be reflected in the emotional space with the continuous mapping function.

Figure 4 may also remind people of the face manifold [28] that means the variations of face images can be represented as low dimensional subspace embedded in the high dimensional image space. Chang et al. [3] learned the structure of the facial expression manifold by Locally Linear Embedding (LLE) and Lipschitz embedding. To acquire the ability of data generalization, Shan et al. [25] learned the appearance manifold of facial expression by supervised Linear Preserving Projections (LPP). However, none of their manifolds was linked with affective properties such as valence or arousal. Another difference is that most of previous work explores the face manifold by utilizing the adjacent structure in original face feature space whereas we use a much simpler method of data fitting. It is worth to point out that the emotional mapping implemented in this paper was based on a simplest (global) linear function. If advanced embedding technology is adopted with a proper supervision, the quality of this mapping would be improved further. That may be a next research direction.

4. USING PROPER FACE FEATURE

In this Section we discuss the representation of face feature used in emotional mapping, i.e., the $x$ in Equation 1. A good feature should be able to effectively capture the variation on face images. We used three kinds of feature to train the emotional mapping and observe the distribution of expressional faces.

4.1 RAW feature

This means the raw image pixels. As the most basic feature, it is always worth to test. To automatically process
Figure 4: The appearance of expressional faces according to the point layout in Figure 3. Each face patch is centered at its emotional coordinates. To avoid overlapping, some faces are not shown when their places have been occupied by early drawn faces.

large amount of image, face region is detected with a cascade Adaboost classifier [16] and face shape is aligned with an alignment algorithm [33] so that variation of translation and scale are removed. The histogram equalization is then applied to improve the image quality. As a result, a face is normalized as an image patch of $48 \times 48$ pixels, forming a 2304-dimensional feature vector.

4.2 LBP feature

Local Binary Patterns (LBP) [21] is a simple feature capable for representing micro-patterns of face images. It has been reported to be particularly effective for facial expression recognition in low-resolution images [26]. In our experiment the face is divided into $7 \times 7$ sub-regions, $\text{LBP}_{5,7}^{2}$, with 59 dimensions is extracted for each sub-region, so a $7 \times 7 \times 59 = 2891$ dimensional feature vector is obtained. Note both RAW and LBP are appearance-based feature. Since all the faces in Cohn-Kanade database are frontal, the trained emotional mapping will not consider the appearance change due to view angle. It would be improper to directly handle non-frontal faces with appearance-based feature. Therefore, the next SHP feature is developed.

4.3 SHP feature

The shape variation on face images can be used for facial expression recognition independently with appearance feature [4, 29]. We design a special shape feature (SHP) that manages to capture the deformation of facial organs across view angle. We select 88 landmark points around mouth, nose, eyes, eyebrows, face contour, and build a PCA-based shape model [18] using Cohn-Kanade database. Note this shape model incorporates factors of facial expression and subject identity, with valid distribution of face in frontal view. When a non-frontal face is presented to the model, the reconstructed face could be in frontal view by restricting coefficients of the shape model within a valid range [5]. Figure 5 shows an example where the face has been transformed to frontal view whilst keeping original expression. The restricted model coefficients just constitute the SHP feature vector. It should be aware that the facial expression will be disturbed more or less while “turning” the face view. A tighter restriction on shape model parameters may turn the face to better frontal view, and meanwhile cause more losses on original expression.

While training the emotional mapping, to prevent from
under-fitting problem (the amount of training samples is less than the dimension number of face feature vector), it could use Principle Component Analysis (PCA) to reduce the dimensionality. Different emotional mappings can be trained by using RAW, LBP or SHP features. For the training set, the Cohn-Kanade database, it is observed that the face distribution made by different emotional mappings looks quite similar. The result shown in Figure 3 and Figure 4 was derived using RAW feature based mapping. However, when dealing with faces outside the training set, the mapping result based on different face feature becomes large. In the case of non-frontal face, the SHP feature may be the only good choice.

5. MAPPING EXPRESSIONAL FACES

Since the emotional mapping was developed in a person-independent manner, arbitrary expressional face can be represented as a point on the valence-arousal plane by using Equation 1. The emotional mapping links low-level face feature with affective dimensions. With the mapping, the emotional content of a large set of face images can be visualized as a distribution of points on the plane. This actually creates an emotional index that enables user to retrieval expressional faces by selecting a specific area on the emotion plane.

We show the case of emotional mapping applied to JAFFE database [20]. This database contains 213 face images of 7 basic emotional expressions posed by 10 Japanese female models. Since the faces were captured in frontal-view, all three face features (RAW, LBP, SHP) can be used. Emotional mapping based on different feature was tested respectively. Two criteria were used to judge the quality of mapping result: (i) faces with similar expression should be mapped adjacently on the emotional plane; (ii) the layout of faces should conform to the meaning of affective dimensions. It is found that the LBP-based mapping produces the best result. To avoid overlapping, part of the mapped faces is shown in Figure 6. Since the mapping result takes real value, the classification-based validation methods, such as “leave-one-out”, seems not very suitable for this propose. So we have not yet made quantitative evaluation so far.

To further test the performance of proposed emotional mapping, we conduct experiment on a video clip extracted from a popular Korean TV series “Da-Chang-Jin”. Differing with JAFFE database where the images were captured under well-controlled condition, the faces in real video are subject to large variation of view angle. In this case the appearance-based feature seems unsuitable, so the SHP feature has to be used in the emotional mapping. It is observed that the actors’ expressional face do distribute over the emotional plane according to the affective factors. However, part of the faces is not mapped reasonably. The current SHP feature looks not effective enough to capture face variation all the time, especially in large face view. Efforts are still needed to improve the SHP feature.

It has been pointed out [14] that it is easy to build an interactive interface that allows the viewer to navigate throughout the video according to affective contents. For example, when the mouse pointer is moved on the emotional plane, a set of nearby points that belong to the same shot or scene is highlighted and a preview of relevant clip is shown aside. If the preview satisfies the viewer, a click on mouse button confirms current selection and the play starts from selection point. Otherwise the viewer may continue browse the video by moving the mouse pointer to other area over the emotional plane.

6. CONCLUSION

In this paper, we proposed a simple method to build a mapping that links low-level face feature to emotional dimensions. The emotional mapping was built by a coarse data labeling on Cohn-Kanade database and a simple linear fitting on labeled data. The mapping can be used for arbitrary person and a group of face images can be represented as a distribution of points in the 2-D emotional space characterized by the arousal and valence. We tested the performance of emotional mapping on JAFFE database as well as on a clip of real video. Our preliminary experiment shows that the affective content of face images can be well distributed in the 2-D emotional space. While building the emotional mapping, we explored several face features to adapt different situation, such as using LBP-based mapping for frontal faces, using SHP feature for non-frontal faces in real video. Proper facial feature may make the emotional mapping more effective, which is a future research direction. With current experimental result, we also discussed the possibility of building an interactive interface based on the emotional plane where user is able to navigate throughout the video according to affective content.
7. REFERENCES


