B

Advances in Emotional Picture Classification

Yu-Jin Zhang
Department of Electronic Engineering, Tsinghua University, China

INTRODUCTION

One picture is worth more than thousand words. A picture contains not only visual features as well as object and scene information, but also emotional meanings that are usually represented in adjective form, for example, happy, sad, fear, *etc*. (Wang, 2008).

In content-based visual information retrieval (CBVIR), emotional-based one is considered as in the highest level, as it involve the subjective perception and understanding of visual information. On the other side, emotion recognition is considered as one of the important part to develop in human-human and human-computer interaction.

While the recognition of emotions revealed by facial expression has been mature enough to develop available emotion recognition software, automatic classification of emotions evoked by visual scenes in pictures is still a quite preliminary and challenging area of research (Li, 2010).

Few works have been conducted in this field. One typical approach is based on holistic image feature (Yanulevskaya, 2008). It decomposed complex scenes according to an annotated object vocabulary, and assigned a similarity score to all words for each region in an image. Different combinations of visual words in a similarity histogram provided a sufficient characterization of a complex scene picture. Wiccest features and Gabor features were then used for regional feature extraction, and SVM framework was adopted for supervised learning of emotion classes.

Two particular new works are presented here to show the recent advances in this field. One is focused on the middle level semantic extraction, in an intention to fill the gap between low-level visual features and high-level emotional concepts to connect human and machine in a uniform framework. Another is to combine low-level visual features and high-level text features to improve the efficiency of classification procedure.

BACKGROUND

To bridge the "semantic gap" between low-level features and high-level emotional concepts, a middle-level is inserted. Unlike the traditional machine learning methods that directly establish the mapping relationship from the low-level features to high-level emotional concepts, a new concept called latent emotional semantic factors is introduced. The unsupervised probabilistic Latent Semantic Analysis (pLSA) model is used to discover the latent emotional semantic factors.

The procedure is depicted in Figure 1. It starts from "feature extraction," via "latent emotional factors," to "emotion classification." They are corresponding to "low picture feature," "middle semantic level," and "high emotion concept," respectively.

Classification with Latent Semantic Factors

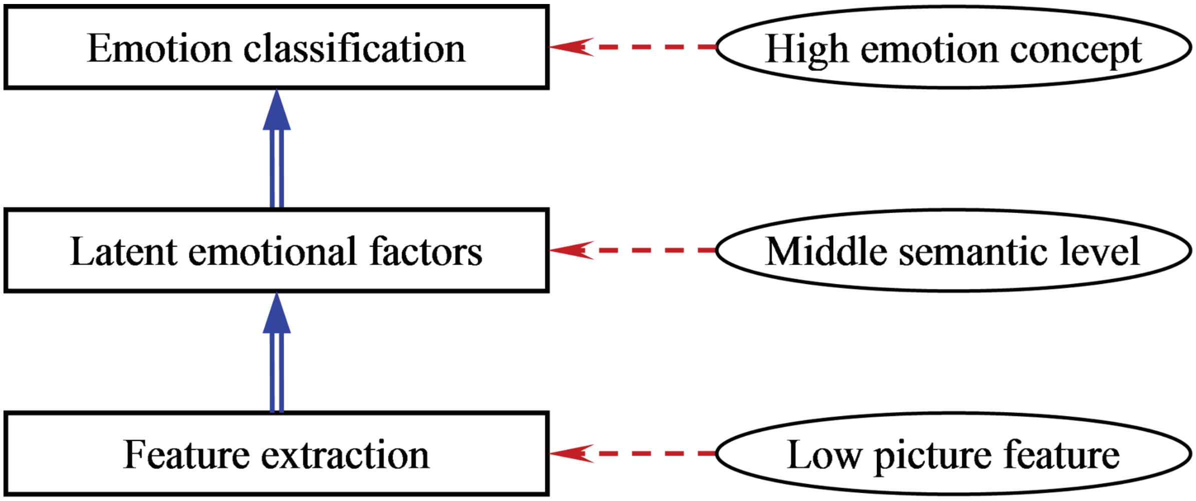
For this procedure, some concepts in different modules are explained as follows.

Low Level Picture Feature: SIFT

To extract low-level picture features, SIFT descriptor (Lowe, 2004) can be used, which could extract distinctive invariant features from images effectively. In practice, SIFT feature is extracted from each subregion of the image, which is densely sampled on an

DOI: 10.4018/978-1-4666-5888-2.ch048

Figure 1. Three levels of emotional picture classification



overlapping regular grid. It is shown that thus extracted features are invariant to image scale and rotation, and insensitive to affine distortion, change in 3D viewpoint, addition of noise and change in illumination (Lowe, 2004; van de Sande, 2010).

Discover Latent Emotional Semantic Factors: pLSA

Probabilistic Latent Semantic Analysis (pLSA) is a generative model, which was originally developed for topic discovery in the text research (Hofmann, 1999; 2001), and then it was applied to discover the latent topic of object categories in scene classification (Bosch, 2006; Sivic, 2005).

It is supposed that there are N documents containing words from a vocabulary of size M. The text document is characterized as an N by M co-occurrence table, where $n(d_i, w_j)$, $i = 1, \ldots, N$, $j = 1, \ldots, M$, denotes how often the word w_j occurs in the document d_i. This is the bag of words model that associates an unobserved class variable $z_k \in \{z_1, \ldots, z_K\}$ with each observation, an observation being the occurrence of a word in a particular document. In order to deduce the pLSA model, some probabilities are defined as follows (Hofmann, 2001): $P(w_j|d_i)$ is used to denote the probability that a word occurrence w_j will be observed in a particular document d_i; $P(w_j|z_k)$ denotes the class-conditional probability of a specific word conditioned on the unobserved class variable z_k; and finally $P(z_k|d_i)$ denotes a document specific probability distribution over the latent variable space. As a result, a joint probability model over $M \times N$ is obtained:

$$P(d_i, w_j) = P(d_i)P(w_j \left| d_i \right.) \tag{1}$$

$$P(w_j \left| d_i \right.) = \sum_{k=1}^{K} P(w_j \left| z_k \right.)P(z_k \left| d_i \right.) \tag{2}$$

Then maximum likelihood estimation is used to maximize T:

$$T = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_j) \tag{3}$$

The standard procedure for maximum likelihood estimation is the Expectation Maximization (EM) algorithm. With it, the conditional distributions $P(w_j|z_k)$ and $P(w_j|d_i)$ can be finally learned.

In the terms of for emotional picture classification, the images are considered as documents (d), and the latent emotional semantic factors are defined as latent topics (z). Firstly, K-means algorithm is used to cluster low-level picture features to form the emotional vocabulary table (w) of size M. Next, the pLSA method is applied to learn the latent emotional semantic factors

Related Content