Multi-Instance Hidden Markov Model For Facial Expression Recognition

Chongliang Wu¹, Shangfei Wang ² and Qiang Ji²
¹ School of Computer Science and Technology, University of Science and Technology of China
² ECSE Department, Rensselaer Polytechnic Institute
Email: {clwzkd@mail.ustc.edu.cn, sfwang@ustc.edu.cn, qji@ecse.rpi.edu}

Abstract—This paper presents a novel method for facial expression recognition using only sequence level labeling. With facial image sequences containing multiple peaks of expression, our method aims to label these sequences and identifies expression peaks automatically. We formulate this weakly labeled expression recognition as a multi-instance learning (MIL) problem. First, image sequences are clustered into multiple segments. After segmentation, image sequences are regarded as bags in MIL and the segments in the bags are viewed as instances. Second, bags data are used to train a discriminative classifier which combined multi-instance learning and discriminative Hidden Markov Model (HMM) learning. In our method, HMM is used to model temporal variation within segments. We conducted experiments on CK+ database and UNBC-McMaster Shoulder Pain Database. Experimental results on both databases show that our method can not only label the sequences effectively, but also locate apex frames of multi-peak sequences. Besides, the experiments demonstrate that our method outperforms state of the art.

I. INTRODUCTION

Facial expression recognition has attracted increasing attention due to its wide applications in human-computer interaction. Current research of facial expression recognition can be divided into two categories: frame-based and sequence based. The former usually recognizes expressions from apex expression frame or every frame. The latter analyzes expressions from facial image sequences. Most research of frame-based expression recognition requires the pre-identified apex facial images and their corresponding expression labels, or labels for each frame, while sequence based work need the pre-segmented image sequences, usually from neutral images to apex images, and sequence level labels. Since the spontaneous facial expression videos could feature multi-peaks, manually identifying the apex frame and segmenting expression sequence from the onset frame to the apex frame are very time consuming and prone to error. Compared with expression annotation with pre-identified apex images or pre-segmented expression sequences, weakly labeled data in the form of sequence level ground-truth without providing the onset, duration, and frequency of facial expressions within the video is much easier to annotate, and can be labeled with great accuracy. Thus, in this paper, we formulate sequence-based expression recognition as a multi-instance learning (MIL) problem, in which image sequences are regarded as bags, and image segments of the sequences are viewed as instances in the bags. Multi-instance learning assumes that every positive bag contains at least one positive instance, and every negative bag contains only negative instances. Instead of an exact label for every instance, multi-instance learning algorithms only require bag-level annotations. Through multi-instance learning, we can obtain the temporal segments and their corresponding expression labels simultaneously with only sequence level ground-truth.

To the best of our knowledge, there is only one work to recognize expression using multi-instance method. Sikka et al. [1] proposed multiple segment multi-instance learning method for pain recognition and localization. First, facial expression sequences are clustered into several segments. Second, features are extracted from every frame within segments and fused into one single feature vector. Third, MilBoost framework is adopted for multi-instance learning. Experiments were conducted on the UNBC-McMaster Shoulder pain database, demonstrating the advantages of there proposed approach. Although the fused feature vector in [1] captures temporal information to a certain extent, the classifier they used is static, and it hence failed to capture the temporal variation in segment, which is crucial for expression analysis. Therefore, in this paper, we propose to use a HMM to model temporal variation within a segment and develop a novel training process to combine discriminative HMM training with MIL.

Although much effort has been made to develop multi-instance learning algorithms, such as diverse density, Citation-kNN, ID3-MI, RIPPER-MI, and BP-MIP [2], only few works consider temporary modeling for multi-instance learning. Yuksel [3] proposed multiple instance hidden Markov model (MI-HMM) for landmine detection. MI-HMM is accomplished via Metropolis-Hastings sampling. Although they introduced HMM in the frame of multi-instance to model the temporal information in time-sequence data, Metropolis-Hastings sampling, which is a Bayesian inference, does not guarantee that we can obtain the correct posterior estimate if the class of distributions they consider do not contain the true generator of the data we are observing. In addition, learning HMM parameters numerically via sampling may not guarantee convergence depending on the number of samples collected. Different from [3], we propose to learn HMM analytically through gradient descent that guarantees the convergence. Furthermore, discriminative
learning typically produces better classification results because the learning and testing criteria are the same.

Thus, in this paper, we formulate expression recognition as a multi-instance learning problem, and propose a discriminative multi-instance HMM learning algorithm, in which HMMs are adopted to capture the dynamic within instances. Experiments on the Extended Cohn-Kanade (CK+) database [4] and the UNBC-McMaster Shoulder Pain Expression Archive Database [5] demonstrate the effectiveness of our proposed approach.

II. MIL–HMM METHOD FOR EXPRESSION RECOGNITION

A. Method Framework

The framework of our approach is shown in Fig. 1, which consists of image sequence segmentation and multi-instance HMM models. First, an image sequencing is clustered into several segments, and the displacements of the facial feature points between current frame and last frame are extracted as the features. After that, discriminative multi-instance HMM is proposed to capture the temporary information in facial segments. We further propose to learn HMM analytically through gradient descent. During testing, we first estimate labels of image segments using the learned HMM, and then infer the label of an image sequence through multi-instance inference. Simultaneously, we obtain the peaks of expression contained in the sequence. The details are described as follows.

B. Facial image sequence segmentation

Each facial image sequence is cut into several segments, which are regarded as instances in a bag. The normalized cuts (Ncut) algorithm [6] is adopted to cluster an image sequence into several segments. The whole sequence is represented as a weighted undirected graph $G = (V, E)$, where $V$ denotes every frame, and $E$ is formed as the links between every pair of nodes. The weight between edges, $W$, is a function of the similarity between nodes $r$ and $s$. The object of Ncuts algorithm is to minimize the following function by removing edges between two nodes,

$$Ncut(V_1, V_2) = \text{cut}(V_1, V_2) + \text{assoc}(V_1, V_2)$$

where $V_1, V_2$ are two disjoint node sets of $G$, and satisfy the constrains of $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$; $\text{cut}(V_1, V_1)$ is the degree of dissimilarity between $V_1, V_2$, and can be computed as total weight of the edges that have been removed; $\text{assoc}(V_1, V)$ is the total connection from nodes in $V_1$ to all nodes in the graph; Similarly, $\text{assoc}(V_2, V)$ denotes the total connection from nodes in $V_2$ to all nodes in the graph.

$$\text{cut}(V_1, V_2) = \sum_{u \in V_1, v \in V_2} W(u,v)$$

$$\text{assoc}(V_1, V) = \sum_{u \in V_1, t \in V} W(u,t)$$

Since segments consist of continuous frames, the weight between frame $f_r$ and frame $f_s$ of sequence $i$, i.e. $W_i(r,s)$, should contain feature similarity as well as the time index of two frames, as shown in Eq. 1:

$$W_i(r,s) = \exp \left( -\frac{\text{dist}(F(f'_r), F(f'_s))}{\delta_f} \right)^2 + \exp \left( -\frac{t_r - t_s}{\delta_t} \right)^2$$

where $F(f)$ indicates the feature vector of frame $f$, the $\text{dist}$ indicates the distance between vectors. Here, we use Euclidean distance, $t$ is the relative time of frame in the sequence. $\delta_f$ and $\delta_t$ are the parameters to regularize the weight matrix.

For the features, we use facial landmarks provided by the databases. First, the face images are normalized according to the binocular position. Then the displacements of feature points between current frame and last frame are used as the features.

C. Multi-instance HMM for expression recognition

Let $D = \{(S, Y)\}$ denote facial image sequences data with only sequence level expression labels, where $S = \{S_1, S_2, ..., S_N\}$ is image sequences set, and $Y = \{Y_1, Y_2, ..., Y_N\}$ is the corresponding sequence-level expression labels set. Each image sequence (i.e a bag) $S_i$ may consist of several segments: $X_i = \{X_{i1}, X_{i2}, ..., X_{in}\}$, which are regarded as instances in the bag. $y_i = \{y_{i1}, y_{i2}, ..., y_{in}\}$ denote the corresponding instance labels. During training, we only have the labels for the whole image sequences. The purpose of multi-instance learning is to maximize conditional log-likelihood (CLL) as shown in Eq. 2.

$$\text{CLL}(\Theta) = \sum_{S_i, Y_i=1} \log P(Y_i=1|S_i, \Theta) + \sum_{S_i, Y_i=0} \log P(Y_i=0|S_i, \Theta)$$

where,

$$P(Y_i=1|S_i) = \max_{j=1}^n \{P(y_{ij} = 1|X_{ij})\}$$

and $P(Y_i = 0|S_i) = 1 - P(Y_i = 1|S_i)$, accordingly. Here, $P(y_{ij} = 1|X_{ij})$ denotes the posterior probability of an instance $X_{ij}$ in $S_i$. In this way, the probability that a bag is positive depends on the maximum probability that an instance in it is positive.

Since max function is not differentiable, Noisy-Or (NOR) model is adopted as an approximation, shown in Eq. 4.

$$P(Y_i=1|S_i) = 1 - \prod_{j=1}^n (1 - P(y_{ij} = 1|X_{ij}))$$
Here, HMM is adopted as the instance classifier $P(y_{ij}|X_{ij})$. As a generative model, HMM provides the joint probability of all observed variables and hidden variables. Thus, the joint probability of observed variables is shown in Eq.5:

$$P(X_{ij}|\Theta) = \sum_{\pi_{ij}} P(\pi_{ij}) \cdot P(X_{ij}|\pi_{ij})$$

$$= \sum_{\pi_{ij}} P(\pi_{1,i;j}) \cdot \prod_{t=1}^{T-1} P(\pi_{t+1,i;j} | \pi_{t,i;j}) \cdot \prod_{t=1}^{T} P(X_{t,i;j}|\pi_{t,i;j})$$

where $X_{ij} = \{X_{1,i;j}, X_{2,i;j}, \ldots, X_{T-1,i;j}, X_{T,i;j}\}$ is the observed variables of the $j$th instance in $i$th bag, $\pi_{ij} = \{\pi_{1,i;j}, \pi_{2,i;j}, \ldots, \pi_{T-1,i;j}, \pi_{T,i;j}\}$ express the hidden variables, and $\Theta$ are the parameters.

For HMM, the local conditional probability distribution $P(X_{t,i;j} | \pi_{t,i;j} = k) \sim N(\mu_{ijk}, \Sigma_{ijk})$. To ensure the covariance matrix is positive semidefinite, $\Sigma_{ijk}$ is parameterized as $A = \Sigma_{ijk}^{\frac{1}{2}}$ [7]. For the transition probability, $\theta_{a|b} = P(\pi_{t+1,i;j} = a | \pi_{t,i;j} = b)$, we introduce a transformed parameter $\beta_{a|b}$ to incorporate the constraints $\theta_{a|b} \geq 0$ and $\sum_{a} \theta_{a|b} = 1$, where $\theta_{a|b} = \sum_{e} e^{\beta_{a|b} e}$. In order to maximize the conditional log-likelihood as shown in Eq.2, we should transfer the log-likelihood of HMM to conditional log-likelihood accord to Bayesian theorem:

$$P(y_{ij} = c | X_{ij})$$

$$= \frac{P(X_{ij} | y_{ij} = c) P(y_{ij} = c)}{\sum_{y_{ij} = 0} P(X_{ij} | y_{ij} = c') P(y_{ij} = c')}$$

(6)

where $c \in \{0, 1\}$ and $c' \in \{0, 1\}$.

Thus, we train HMM discriminatively for the multi-instance learning using gradient descent as shown in Eq.7

$$\frac{\partial LL(\Theta)}{\partial \Theta} \approx \sum_{S_i, Y_i=1}^{n} \sum_{j=1}^{n} P(Y_i = 0 | S_i) \cdot P(y_{ij} = 1 | X_{ij}) \cdot \frac{\partial \log P(y_{ij} = 1 | X_{ij})}{\partial \Theta}$$

(7)

$$- \sum_{S_i, Y_i=0}^{n} \sum_{j=1}^{n} \frac{\partial \log P(y_{ij} = 0 | X_{ij})}{\partial \Theta}$$

The right hand side of Eq.7, the derivative, $\frac{\partial \log P(y_{ij} = 1 | X_{ij})}{\partial \Theta}$, and $\frac{\partial \log P(y_{ij} = 0 | X_{ij})}{\partial \Theta}$, can be translated into $\frac{\partial \log P(X_{ij}|y_{ij}=1)}{\partial \Theta}$ and $\frac{\partial \log P(X_{ij}|y_{ij}=0)}{\partial \Theta}$ by applying Bayesian theorem as Eq.6 [7], and can be expressed as follows:

$$\frac{\partial \log P(y_{ij} = c | X_{ij})}{\partial \Theta_{c'}} = \begin{cases} [1 - P(y_{ij} = c | X_{ij})] & c' = c \\ -P(y_{ij} = c' | X_{ij}) & \frac{\partial \log P(X_{ij}|y_{ij}=c')}{\partial \Theta_{c'}} \neq 0 \\ \frac{\partial \log P(X_{ij}|y_{ij}=c)}{\partial \Theta_{c'}} & c' \neq c \end{cases}$$

(8)

where $c \in \{0, 1\}$ and $c' \in \{0, 1\}$.

Therefore, we obtain the derivative of $P(X_{ij})$ to parameters $\mu_{ijk}$, $A_{ijk}$ and transition probability $\beta$ as shown in Eq.9, Eq.10 and Eq.11 respectively.

$$\frac{\partial \log P(X_{ij})}{\partial \mu_{ijk}} = \sum_{t=1}^{T} P(\pi_{t,i;j} = k | X_{ij}) A_{ijk}^2 (X_{t,i;j} - \mu_{ijk})$$

(9)

$$\frac{\partial \log P(X_{ij})}{\partial A_{ijk}} = \sum_{t=1}^{T} P(\pi_{t,i;j} = k | X_{ij})$$

$$\left[ A_{ijk}^{-1} - A_{ijk} (X_{t,i;j} - \mu_{ijk}) (X_{t,i;j} - \mu_{ijk})^T \right]$$

(10)

where $P(\pi_{t,i;j} = k | X_{ij})$ can be computed using the forward algorithm for HMM.
The detailed of parameter learning algorithm is listed in Algorithm_1.

D. Testing

During testing, an image sequence $S_i$ is first clustered into several segments $(X_{i1}, ..., X_{in})$ using Ncut algorithm described in Section II-B After that, the likelihood probability of every segment $P(X_{ij}|y_{ij})$ is obtained using the learned HMMs according to Eq.5. Thus, we obtain the expression category of every segments according to Eq.12

$$y_{ij} = \{ P(X_{ij}|y_{ij} = 1) > P(X_{ij}|y_{ij} = 0) \}$$

Then, the label of an image sequence is inferred using Eq.3 or Eq.4.

III. EXPERIMENTS AND RESULTS

A. Experimental conditions

We conducted multi-instance expression recognition on two benchmark databases, i.e. the Extended Cohn-Kanade (CK+) database [4] and the UNBC-McMaster Shoulder Pain Expression Archive Database [5].

The CK+ database is a posed expression database, which consists of 327 facial expression image sequences, from neutral frame to apex frame. The length of sequences ranges over 5 to 70, and each expression sequence belongs to one of seven expression categories (i.e. angry, contempt, disgust, fear, happiness, sadness, surprise). Table.I shows the distribution of image sequences over seven expressions. For this database, we generate negative sequences using neutral frame from the corresponding positive sequence by simply duplication. negative sequences contains only neutral frames.

The UNBC-McMaster Shoulder Pain database is a spontaneous pain expression database, including 200 image sequences from 25 subjects. The length of the sequences vary from $60 \sim 600$. The Observed Pain Intensity (OPI), ranging from 0 (no-pain observed) to 5 (extreme pain observed), are assigned to each frame of all sequences by an independent observer. In our experiments, the sequences with $OPI \geq 3$ are labeled as positive (‘pain’) bags, and the sequences with $OPI = 0$ are labeled as negative (‘no pain’) bags[1]. Thus, we obtain 57 positive bags and 92 negative bags.

Both databases provide feature points for each frame, which are used as features in our experiments. Leave-one-subject-out cross validation method is adopted.

B. Experimental Result and Analysis On CK+ Database

The expression recognition results on CK+ database are shown in Table.II. From Table.II, we can find that the average accuracy is $98.54\%$. Furthermore, the performance of anger and sadness recognition reaches 100\%. This demon-

C. Experimental Result and Analysis On UNBC-McMaster Shoulder Pain database

<table>
<thead>
<tr>
<th>Expression</th>
<th>Accuracy</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>100%</td>
<td>1</td>
</tr>
<tr>
<td>contempt</td>
<td>98.15%</td>
<td>0.9714</td>
</tr>
<tr>
<td>disgust</td>
<td>97.18%</td>
<td>0.9558</td>
</tr>
<tr>
<td>fear</td>
<td>97.35%</td>
<td>0.9283</td>
</tr>
<tr>
<td>happiness</td>
<td>99.52%</td>
<td>0.9927</td>
</tr>
<tr>
<td>sadness</td>
<td>100%</td>
<td>1</td>
</tr>
<tr>
<td>surprise</td>
<td>97.59%</td>
<td>0.9652</td>
</tr>
<tr>
<td>average</td>
<td>98.54%</td>
<td>0.9776</td>
</tr>
</tbody>
</table>

The confusion matrix is shown in Table.III. From Table.III, we can find the accuracy is 85.23\% and the F1-score is 0.78.

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
<th>classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive</td>
</tr>
<tr>
<td>positive</td>
<td>5</td>
</tr>
<tr>
<td>negative</td>
<td>18</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>F1-score</td>
<td></td>
</tr>
</tbody>
</table>
TABLE IV
COMPARISON BETWEEN OUR METHOD AND OTHER PAIN DETECTION ALGORITHMS

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th># subjects</th>
<th># sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIL-HMM</td>
<td>85.23%</td>
<td>25</td>
<td>149</td>
</tr>
<tr>
<td>MS-MIL (K. Sikka et.al)</td>
<td>83.7%</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>P. Lucey et.al [9]</td>
<td>80.99%</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>A.B. Ashraf et.al [10]</td>
<td>68.31%</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>A.B. Ashraf et.al [10]</td>
<td>81.21%</td>
<td>21</td>
<td>84</td>
</tr>
</tbody>
</table>

Fig. 2. Expression localization examples on CK+ database

strates the effectiveness of our proposed approach.

In addition to expression recognition, our method can locate the expression segments contained in the positive image sequences. The posterior probability of segments indicate the score of the segments being positive. Since the segmentation process clusters similar frames together, the score of each frame of one segment is the same as that of the segment. Fig.2 shows two samples obtained by our multi-instance HMM. According to the criterion we used for image sequence extension of the CK+ database, each positive sequence has only one positive instance in the middle. From these two samples, we can find that the inferred scores are consistent with the ground truth.

C. Experimental Result and Analysis On UNBC-McMaster Shoulder pain database

Expression recognition results on the UNBC-McMaster Shoulder pain database are shown in Table.III. From the table, we can find that our approach achieves 85.23% of accuracy and 0.78 of F1-score, demonstrating its effectiveness. From the confusion matrix, we may find that it is easier to recognize negative sequences than positive ones. This is probably due to the unbalanced data.

Similar as CK+ database, two examples are shown in Fig.3 to demonstrate the expression localization capability of our method. The green line is the normalized PSPI value, which is a frame level ground truth.

From Fig.3(a), we can find that our method can locate multi-peak pain segments, such as frames around 80 and 280. From Fig.3(b), we find that the posterior probability we predicted can locate some ambiguity segments correctly, such as frames around 175 and 450.

D. Comparison with related work

Till now, there is only one work that recognized expression using multi-instance learning, and conducted experiments on the UNBC-McMaster Shoulder pain database. Therefore, we only compare our work with related work on the UNBC-McMaster Shoulder pain database as shown in Table .IV.

From the table, we can find that both our approach and Sikka et.al’s outperform [9] and [10]. These two works formulated pain recognition as a traditional machine learning problem, and used frame-based expression recognition method, support vector machines (SVM) are adopted as classifier to assign labels for each frame. It demonstrate the superior of multi-instance learning for facial expression recognition with weakly labeled data. Compared with Sikka et.al’s approach, our method achieves better performance. Although Sikka et.al also formulates expression recognition as a multi-instance learning problem, the classifier they used is static, which can not fully capture the dynamic embedded in expression sequences. While our approach adopts HMM to model the temporary variation within segments.

IV. CONCLUSION

In this paper, a method for expression recognition using only weakly labeled image sequencing data by applying MIL framework in combination with Discriminative HMM learning approach is proposed. We first demonstrate that
how multi-instance be created. We divide facial every image sequence into several segments using Ncut clustering algorithm. In this way, image sequences can be regarded as bags and segments in the bags are viewed as instances. Then, the detail of MIL-HMM method for expression recognition is presented. We trained a classifier under MIL framework for expression recognition. In order to better capture temporal variation contained in segments, HMM is applied. During training, we learn the parameters by combining multi-instance learning and discriminative HMM training algorithm. We experiment on CK+ database and UNBC-McMaster Shoulder Pain Database. Experimental results on both databases show that our method can not only label the sequences effectively, but also locate apex frames of multi-peak sequences. Further more, the results demonstrate that our method outperforms both traditional machine learning and other multi-instance learning based methods.

REFERENCES