The detection of concept frames using Clustering Multi-Instance Learning

D.M.J. Tax, E. Hendriks
Pattern Recognition Lab
Delft University of Technology
{*D.M.J.Tax,E.A.Hendriks*}*@tudelft.nl*

M.F. Valstar, M.Pantic
Department of Computing
Imperial College London
{*Michel.Valstar,m.pantic*}*@imperial.ac.uk*

***Abstract*—The classification of sequences requires the combination of information from different time points. In this paper the detection of facial expressions is considered. Experiments on the detection of certain facial muscle activations in videos show that it is not always required to model the sequences fully, but that the presence of specific frames (the concept frame) can be sufficient for a reliable detection of certain facial expression classes. For the detection of these concept frames a standard classifier is often sufficient, although a more advanced clustering approach performs better in some cases.**

***Keywords*-classification, multi-instance learning, time series classification**

I. Introduction

For the automatic analysis of human behavior in video, the detection and classification of emotions are important elements. One modality that can be used to detect and classify emotions are the facial expressions. Facial expressions are often described using the Facial Action Coding System in terms of Action Units (AUs) [1]. AUs encode the activation or relaxation of all facial muscles involved in facial expression. This allows the objective encoding of any facial expression, not just expressions of emotion. The sequence of activations and relaxations can be used to classify emotions [2].

In greater detail, an AU activation consists of a number of temporal phases. Starting from a neutral state, an AU will go through an onset, an apex (peak) and an offset phase, before returning to the neutral state again. Although the timing and the (relative) durations of the AU activations are very important for the classification of facial expressions, the first step is the detection of the AU activations themselves in the video stream. The question is if the detection of the presence of an AU requires the modeling of the full time series or if the presence of a single key frame may be sufficient. A single-frame detector may be faster and more simple to implement, with the potential drawback that important temporal information is lost. This paper investigates the two different approaches to classify sequences of events.

For the detection of AU activations either frame-by-frame detectors have been proposed, or detectors using pre-segmented sequences. See [3] for an overview of AU detection methods. In pattern recognition objects, or events, are encoded by feature vectors of fixed length, and are assigned to a single class [4]. In many applications this reduction to a single feature vector is very difficult, sometimes even impossible. This happens not only in the classification of time series of variable length, but also in the description of images, texts, or complex physical objects. In these applications Multi-Instance Learning (MIL) can be considered [5]. MIL represents an object or event, by a *bag* of feature vectors. These feature vectors are often assumed to be independent. Furthermore, the individual feature vectors are not labeled, but the complete bags are. In the original formulation a bag is considered positive when at least one vector is member of a so-called *concept*, and it is considered negative when none of the vectors is member of the concept. In later literature, the constraint of a single positive vector is often relaxed, and sometimes a combining rule is learned [6].

A standard way of learning a label sequence from a time series is by using Conditional Random Fields (CRFs) [7]. CRFs are graphical models that assume a sequential structure of the data. The CRF estimates the conditional probability of the states Y given the observations X, $P(Y|X)$, and do not model the full probability density of the observations and labels $P(X,Y)$, they tend to be more flexible and easier to optimize than Hidden Markov Models [8]. On the other hand, during training they require a fully labeled training set where for each time point a label is provided.

In section II, we first discuss the two approaches to classify sequences, the Multi-Instance approach where concept frames are learned, and the approach to model sequences using CRFs. We also introduce a simple Multi-Instance learner that optimizes a concept frame on the training data. In section III, experiments on movies containing faces are presented where both approaches are applied and results are discussed. Finally, in section IV, we finish with some conclusions.

II. Theory

Assume we are given a collection of M labeled sequences (or bags): $\{(X_i, y_i),\ i = 1, ..., M\}$ with binary labels $y_i \in \{-1, +1\}$. Each bag X_i contains a variable number T_i of instances $X_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT_i}\}$, where $\mathbf{x} \in \mathbb{R}^d$ are feature vectors in a d dimensional feature space. Each of the

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.715

instance vectors may also be labeled, although in this paper it is assumed that these labels are unknown. Furthermore, it is unclear how the bag label y_i is derived from the instance labels, if they would have been known.

A. Multi-instance learning

A straightforward approach to classify a bag is to train a standard classifier f on the individual feature vectors $\mathbf{x}_{ij}$ using the bag label y_i as labels for the instances. Classifier f is assumed to generate a positive score for each input vector, i.e. the larger $f(\mathbf{x})$, the more certain classifier f is that object $\mathbf{x}$ belongs to the positive class. Then a combining rule h has to integrate the individual outcomes to a bag outcome:

$$\hat{y}(X_i) = h(f(\mathbf{x}_{i1}), ..., f(\mathbf{x}_{iT_i})). \tag{1}$$

This bag outcome $\hat{y}(X_i)$ is thresholded to decide if the bag is labeled positive or negative.

One possibility for this combining rule h is a maximum-rule that selects the most positive output over all feature vectors in the bag. This is consistent with the original idea of Multi-Instance learning: a bag is labeled positive, when at least one instance is labeled positive. This approach is outlier sensitive though, and a more robust alternative for h is the 'quantile rule'. Instead of selecting the absolute maximum positive output, all outputs are ordered $f(\mathbf{x}_{i1}) > f(\mathbf{x}_{i2}) > ... > f(\mathbf{x}_{iT_i})$, and the q-th quantile value is returned:

$$h_q(f(\mathbf{x}_{i1}), ..., f(\mathbf{x}_{iT_i})) = f(\mathbf{x}_{i\lfloor qT_i+0.5\rfloor}), \tag{2}$$

where $\lfloor qT + 0.5\rfloor$ rounds the value qT to the nearest integer value. For instance, $q = 0.5$ returns the median output value, while for $q = 1/T_i$ the maximum value is obtained. Depending on the value of q, the quantile rule can be made insensitive to outlier outputs of some classifiers. Note that in this approach where individual classifier outputs f_i are combined, the classifier f has to be trained on labeled instances. Because the instance labels are copied from the bag labels, and some of the instances in a bag actually originate from the negative (background) class, it may be expected that the final classifier is not optimal.

In this paper the clustering Multi-instance learner is constructed to exploit the original MIL assumption that a bag is labeled positive when at least one instance vector is member of a so-called concept. The concept is modeled by a spherical area in feature space, parametrized by a center and a radius. The center location of this area is selected from a collection of locations that is obtained by some clustering procedure on all instances of the positive bags (in this paper we used the standard k-means clustering). The distance to the concept center is used as the instance classifier f, and the quantile rule is then applied. Each cluster is tested subsequently, and the cluster with the best performance on the training set (in this paper, in terms of the Area under the Receiver Operating Characteristic curve, AUC [9]) is selected. The full training procedure therefore looks like:

1) Assume the number K of potential concept centers, and a quantile level $0 < q < 1$.
2) Cluster the instance vectors from all positive bags $\{X_i : y_i = +1\}$ into K clusters, and obtain cluster centers $\mathbf{c}_k, k = 1, .., K$.
3) Compute for all clusters and all instances in all bags the classifier output:
$$f^k(\mathbf{x}_{ij}) = \exp(-\|\mathbf{x}_{ij} - \mathbf{c}_k\|^2) \tag{3}$$
4) Compute for all clusters k and all bags i the bag output $\hat{y}(X_i)$, using equation (1) and equation (2).
5) Compute for all training clusters the AUC using the bag output, and select the cluster center $\mathbf{c}_*$ with the highest AUC.
6) If needed, the decision threshold on the bag output $\hat{y}(X_i)$ is found by minimizing the total classification error.

For the evaluation of a new bag X, the classifier output (3) for each instance vector in the bag has to be computed, and these outputs have to be combined using (1).

Similar approaches have been proposed in literature. Often they are inspired by the idea of modeling the 'positive concept'. One approach is the Maximum Diverse Density [10]. It optimizes the target concept position $\mathbf{c}$ by maximizing the so-called diverse density:

$$\arg\max_{\mathbf{c}} \prod_{i:y_i=+1} Pr^+(\mathbf{c}|X_i) \prod_{i:y_i=-1} Pr^-(\mathbf{c}|X_i), \tag{4}$$

where $Pr^+(\mathbf{c}|X_i)$ is the probability that at least one vector from a positive bag did not miss the target concept. This can be computed like $Pr^+(\mathbf{c}|X_i) = 1 - \prod_j(1 - P(\mathbf{c}|\mathbf{x}_{ij}))$. Furthermore, the probability that a vector matches the concept is modeled by a circular Gaussian-shaped distribution $P(\mathbf{c}|\mathbf{x}_{ij}) = \exp(-\|\mathbf{x}_{ij} - \mathbf{c}\|^2)$ analogous to (3). Similarly, all vectors in the negative bags should miss the concept, which can be computed with: $Pr^-(\mathbf{c}|X_i) = \prod_j(1 - P(\mathbf{c}|\mathbf{x}_{ij}))$. Unfortunately, the optimization of (4) is very complicated. It requires a careful gradient ascent optimization, with high risks of obtaining a poor local optimum. The clustering Multi-Instance learning formulation is actually an approximation to the maximum diverse density formulation, except that the concept position $\mathbf{c}_*$ is constrained to be on the cluster centers, fitted on the positive training samples. When a simple clustering method like k-means clustering is used, the clustering MIL becomes several orders faster than the Diverse Density approach.

B. Modeling the time series

For the situation in which the instance vectors in bag X_i constitute a (time) sequence, explicit sequence models can be used. A model that uses the label information during training, is the Conditional Random Field, CRF [7]. It is a chain-shaped graphical model, that predicts a full label sequence over all time points. The posterior probability of each of the possible label sequences is conditioned on (in principle) the complete sequence of feature vectors. To simplify the model, we assume that the posterior probability of a label y_t at time t is only conditionally dependent on the feature vector $\mathbf{x}_t$ for that time point, and the prediction of the previous time point label. The conditional dependence of the label on the observations $\mathbf{x}_t$ is modeled by:

$$p(y_t|\mathbf{x}_t) = \frac{1}{Z}\exp(-\mathbf{w}^T\mathbf{x}_t), \tag{5}$$

where $\mathbf{w}$ is a weight vector that is optimized in training and $Z(\mathbf{w})$ is a normalization constant such that $p(y_t|\mathbf{x}_t)$ integrates to 1 (which means a summation over all label sequences). The conditional probability of a label y_t at time point t given a label y_{t-1} at time $t-1$ is directly estimated from the training set and stored in a transition probability matrix $\hat{P}(y_{t-1}, y_t)$. To optimize the free parameters $\mathbf{w}$ on a set of training bags, the conditional log-likelihood $\sum_{m=1}^{M}\log p(\mathbf{y}_m|X_m)$ is optimized, where $p(\mathbf{y}|X) = \prod_{t=1}^{T}\hat{P}(y_{t-1}, y_t)p(y_t|\mathbf{x}_t)/Z$, For more information, see [7].

An alternative sequence model is the Hidden Markov Model (HMM) [8]. This is a chain-shaped graphical model like the CRF, but it is a generative model instead of a discriminative model. Instead of estimating the posterior class probabilities $p(y_t|\mathbf{x}_t)$, it estimates the full probability density of the observed feature vectors $p(X) = p(\{\mathbf{x}_1, ...\mathbf{x}_T\})$, conditioned on the (hidden) state labels. In the HMM the (hidden) label sequence is not constrained, but it can have any value in order to describe the observed data as well as possible. For real-valued observations the conditional probability of an observation $\mathbf{x}_t$ given the state label y_t is often modeled by a normal distribution.

III. Experiments

We consider the detection of activation of face muscles. For this, 211 movie sequences taken from the MMI Facial Expression Database [11] have been AU-annotated, containing in total 19004 video frames. Each sequence shows a certain (posed) facial expression. Originally, for each frame in each movie the activations ('neutral', 'onset', 'apex', 'offset') of 28 Action Units are labeled, but in this paper only the presence of an active AU is of interest. Sequences that have some time points labeled 'onset', 'apex' or 'offset' will be called 'positive', and all other sequences are 'negative'. In the human face the positions of 20 key points are tracked, resulting in a 40 dimensional feature vector $\mathbf{x}_{it} \in \mathbb{R}^{40}$ for each sequence X_i at each time point t. In order to obtain features that are invariant to rigid head motions within one image sequence we intra-register all frames within one sequence by subtracting from $\mathbf{x}_{it}$ the mean value of so-called stable points (i.e. points that by definition only move due to rigid head motion, such as the tip of the nose and the inner eye corners). Variations in size and shape of the face between subjects are minimised by applying a scaling transformation to $\mathbf{x}_{it}$ which again is based on all stable facial points.

Due to space constraints the results for only a few AUs are shown in Table I. Only AUs are used for which the number of positive sequences was around 20 or more and for which different characteristics of the classifiers can be observed. On these datasets 17 classifiers are trained. The first 12 classifiers optimize a concept frame, the second five model the full time sequence. First a simple Fisher classifier [4] is trained on all instance vectors (using the bag labels as instance labels), and combined using the quantile rule h_Q using $q = 0.1$, $q = 0.5$ and $q = 0.9$ or the maximum rule. The results are shown in the first four lines of Table I. The results obtained by this simple classifier reveal very different characteristics for different Action Units. For AU01 and AU06 the maximum rule performs best, but for the other AUs this rule is too noise sensitive. For AU04 or AU12 it is to be preferred to use a 0.1-quantile combination rule. In some situations it also depends on the base classifier f, and better performance can be obtained when the Fisher classifier is replaced by the logistic classifier [12] (like in AU05 or AU09).

In the next six lines of Table I the results for some Multi-Instance learners are shown. The Diverse Density (with a varying number of random initializations k), and the clustering MIL (also with a varying number of clusters, and varying quantile levels q). In some situations they perform very poorly, but the clustering MIL shows competitive performance for the AU05 and AU07.

In the last four lines of Table I the performance of the methods that model the time sequence explicitly are shown. First Hidden Markov Models with a varying number of states N are given (where each state models the emission probabilities using a Gaussian distribution with a full covariance matrix). A random initialization of the HMM parameters is used. Second a CRF with a random initialization or an initialization where $\mathbf{w}$ is initialized with a logistic classifier. It appeared that the random initialization is very unstable, and often a local optimum was found in the optimization. For some detection problems, such as for AU09 and AU11, the time sequence information appears to be very important, and the CRF often outperforms the methods that use only single timepoints.

IV. Conclusions

This paper shows that for the classification of AU sequence data it is not always required to model the full sequence, but that the presence of a concept frame may be sufficient for good classification. The modeling of the sequence requires a suitable model, and sufficient training data to fit the model. But for relatively complex and noisy real world data these two requirements may be too strong. It appears that the detection of a single concept frame can often be done more reliably. The concept frame can

Table I

AREA UNDER THE ROC CURVE PERFORMANCE OF THE CLASSIFIERS ($\times 100$), USING THREE TIMES FIVE-FOLD CROSSVALIDATION. THE BEST PERFORMANCE IS INDICATED IN BOLD, TOGETHER WITH THE PERFORMANCES THAT ARE NOT SIGNIFICANTLY WORSE (USING A ONE-SIDED T-TEST WITH A CONFIDENCE LEVEL OF 5%). BETWEEN BRACKETS THE STANDARD DEVIATION OVER THE THREE RUNS IS SHOWN.

classifier	AU01	AU04	AU05	AU06	AU07	AU09	AU11	AU12
Fisher max	**84.0 (0.0)**	80.2 (0.0)	79.6 (0.0)	**83.7 (0.0)**	73.1 (0.0)	70.0 (0.0)	66.0 (0.0)	81.9 (0.0)
Fisher $q = 0.1$	83.0 (0.0)	**82.4 (0.0)**	80.3 (0.0)	80.2 (0.0)	**76.2 (0.0)**	73.2 (0.0)	66.7 (0.0)	**85.0 (0.0)**
Fisher $q = 0.5$	79.7 (0.0)	76.8 (0.0)	76.8 (0.0)	69.2 (0.0)	75.0 (0.0)	69.1 (0.0)	62.4 (0.0)	77.2 (0.0)
Fisher $q = 0.9$	74.3 (0.0)	70.3 (0.0)	69.9 (0.0)	56.5 (0.0)	74.2 (0.0)	63.1 (0.0)	59.8 (0.0)	62.1 (0.0)
Logistic max	77.1 (0.0)	67.5 (0.0)	**83.7 (0.0)**	73.3 (0.0)	71.8 (0.0)	75.8 (0.0)	82.1 (0.0)	77.5 (0.0)
Logistic $q = 0.1$	75.5 (0.0)	68.2 (0.0)	82.6 (0.0)	72.5 (0.0)	73.4 (0.0)	**78.3 (0.0)**	82.1 (0.0)	80.9 (0.0)
Diverse Dens. $k = 20$	73.9 (3.7)	68.4 (1.7)	82.1 (0.0)	74.6 (1.7)	67.1 (0.1)	75.9 (1.3)	72.3 (0.0)	79.7 (0.3)
Diverse Dens. $k = 50$	74.3 (1.4)	69.5 (1.1)	82.1 (0.0)	74.6 (0.6)	67.1 (0.5)	74.3 (0.0)	72.8 (0.0)	80.2 (1.2)
Clust $K = 20$ max	71.8 (3.3)	68.0 (0.6)	**80.8 (3.2)**	64.7 (1.3)	**74.1 (1.5)**	27.7 (3.0)	71.7 (0.7)	50.7 (2.3)
Clust $K = 20q = 0.1$	68.4 (0.7)	67.0 (1.5)	76.5 (1.6)	63.0 (5.6)	**75.0 (3.2)**	29.3 (2.2)	68.9 (1.5)	43.5 (1.8)
Clust $K = 50$ max	74.3 (1.5)	68.0 (0.3)	**83.2 (0.9)**	61.5 (4.0)	74.6 (1.2)	28.9 (3.2)	73.3 (1.7)	49.2 (2.7)
Clust $K = 50q = 0.1$	72.9 (1.7)	67.0 (1.0)	77.1 (0.3)	68.7 (1.0)	**77.7 (1.2)**	30.5 (0.8)	71.3 (1.3)	45.8 (3.7)
HMM $N = 2$	54.3 (0.5)	57.4 (0.8)	77.1 (0.5)	47.7 (0.5)	66.6 (0.5)	58.1 (0.6)	79.5 (0.7)	38.7 (0.2)
HMM $N = 3$	54.4 (0.6)	56.8 (0.8)	77.8 (0.8)	48.3 (0.9)	66.2 (1.4)	56.0 (1.9)	79.5 (1.4)	38.9 (0.7)
HMM $N = 4$	53.1 (1.6)	56.4 (0.7)	77.0 (0.1)	48.1 (0.4)	67.8 (1.4)	56.5 (1.7)	80.0 (0.1)	39.1 (0.5)
CRF random init.	69.5 (2.4)	65.8 (0.2)	77.2 (0.0)	67.1 (0.8)	73.7 (0.9)	76.8 (0.9)	81.2 (0.0)	80.7 (0.5)
CRF logistic init.	71.3 (2.0)	65.6 (0.6)	76.8 (0.0)	66.7 (0.3)	**74.8 (2.2)**	**78.3 (2.3)**	**88.1 (0.0)**	77.8 (0.4)

be found by training standard classifiers on the individual time frames, and combining the results of the frames in a robust manner. Depending on the classification problem, the robustness or sensitiveness to outliers has to be adjusted. In classification problems that are considered in this paper, this naive approach performed very well. An alternative approach is to use Multi-Instance learning. This approach exploits the fact that the frames are part of a collection (called a 'bag'), but it assumes that the frames are independent and therefore it ignores the sequential nature of the data. Still, for some sequence classification problems it shows very promising performance.

ACKNOWLEDGMENT

This work has been funded in part by the European Communitys 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet). The work of Michel Valstar is further funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE).

REFERENCES

[1] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[2] M. Valstar, "Timing is everything: A spatio-temporal approach to the analysis of facial actions," Ph.D. dissertation, Imperial College of Science, Technology and Medicine Department of Computing, 2008.

[3] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," in *Face Recognition*. I-Tech Education and Publishing, 2007.

[4] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, 2001.

[5] T. Dietterich, R. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[6] H.-Y. Wang, Q. Yang, and H. Zha, "Adaptive p-posterior mixture-model kernels for multiple instance learning," in *Proc. 25th Int'l Conf. Machine learning*, 2008, pp. 1136–1143.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc.18th Int'l Conf. Machine Learning (ICML-2001)*, 2001.

[8] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] A. Bradley, "The use of the Area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[10] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*, vol. 10. MIT Press, 1998, pp. 570–576.

[11] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *IEEE Int'll Conf. Multimedia and Expo*, p. 5, 2005.

[12] J. Anderson, "Logistic discrimination," in *Classification, Pattern Recognition and Reduction of Dimensionality*, ser. Handbook of Statistics, P. Kirshnaiah and L. Kanal, Eds. Amsterdam: North Holland, 1982, vol. 2, pp. 169–191.