LOMo: Latent Ordinal Model for Facial Analysis in Videos

Karan Sikka[1,*]   Gaurav Sharma[2,3,†]   Marian Bartlett[1,*,‡]

[1]UCSD, USA   [2]MPI for Informatics, Germany   [3]IIT Kanpur, India

Abstract

We study the problem of facial analysis in videos. We propose a novel weakly supervised learning method that models the video event (expression, pain etc.) as a sequence of automatically mined, discriminative sub-events (e.g. onset and offset phase for smile, brow lower and cheek raise for pain). The proposed model is inspired by the recent works on Multiple Instance Learning and latent SVM/HCRF – it extends such frameworks to model the ordinal or temporal aspect in the videos, approximately. We obtain consistent improvements over relevant competitive baselines on four challenging and publicly available video based facial analysis datasets for prediction of expression, clinical pain and intent in dyadic conversations. In combination with complimentary features, we report state-of-the-art results on these datasets.

1. Introduction

Facial analysis is an important area of computer vision. The representative problems include face (identity) recognition [52], identity based face pair matching [10], age estimation [1], kinship verification [23], emotion prediction [6], [11], among others. Facial analysis finds important and relevant real world appli-
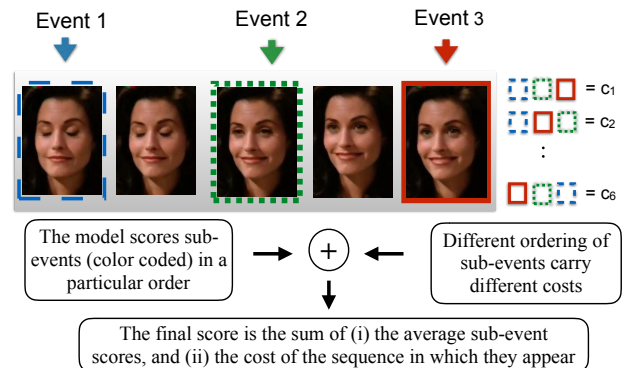

Figure 1: Illustration of the proposed approach.

cations such as human computer interaction, personal robotics, and patient care in hospitals [37, 25, 42, 5]. While we work with videos of faces, i.e. we assume that face detection has been done reliably, we note that the problem is pretty challenging due to variations in human faces, articulations, lighting conditions, poses, video artifacts such as blur etc. Moreover, we work in a weakly supervised setting, where only video level annotations are available and there are no annotations for individual video frames.

In weakly supervised setting, Multiple Instance Learning (MIL) [2] methods are one of the popular approaches and have been applied to the task of facial video analysis [37, 33, 45] with video level, and not frame level, annotations. However, the main drawbacks of most of such approaches are that (i) they use the maximum scoring vector to make the prediction [2], and (ii) the temporal/ordinal information is always lost completely. While, in the recent work by Li and Vasconcelos [17], MIL framework has been extended to consider multiple top scoring vectors, the temporal order is still not incorporated. In the present paper we propose a novel method that (i) works with weakly su-

pervised data, (ii) mines out the prototypical and discriminative set of vectors required for the task, and (iii) learns constraints on the temporal order of such vectors. We show how modelling multiple vectors instead of the maximum one, while simultaneously considering their ordering, leads to improvements in performance.

The proposed model belongs to the family of models with structured latent variables e.g. Deformable Part Models (DPM) [7] and Hidden Conditional Random Fields (HCRF) [43]. In DPM, Felzenszwalb et al. [7] constrain the location of the parts (latent variables) to be around fixed anchor points with penalty for deviation while Wang and Mori [43] impose a tree structure on the human parts (latent variables) in their HCRF based formulation. In contrast, we are not interested in constraining our latent variables based on fixed anchors [7] or distance (or correlation) among themselves [43, 32], but are only interested in modeling the order in which they appear. Thus, the model is stronger than models without any structure while being weaker that models with more strict structure [7, 43].

The current model is also reminiscent of Actom Sequence Model (ASM) of Gaidon et al. [8], where a temporally ordered sequence of sub-events are used to perform action recognition in videos. However, ASM requires annotation of such sub-events in the videos; the proposed model aims to find such sub-events automatically. While ASM places absolute temporal localization constraints on the sub-events, the proposed model only cares about the order in which such sub-events occur. One advantage of doing so is the flexibility of sharing appearances for two sub-events, especially when they are automatically mined. As an example, the facial expression may start, as well as end, with a neutral face. In such case, if the sub-event (neutral face) is tied to a temporal location we will need two redundant (in appearance) sub-events i.e. one at the beginning and one at the end. While, here such sub-events will merge to a single appearance model, with the symmetry encoded with similar cost for the two ordering of such sub-event, keeping the rest same.

In summary, we make the following contributions. (i) We propose a novel (loosely) structured latent variable model, which we call Latent Ordinal Model (LOMo). It mines prototypical sub-events and learns a prior, in the form of a cost function, on the ordering of

such sub-events automatically with weakly supervised data. (ii) We propose a max-margin hinge loss minimization objective, to learn the model and design an efficient stochastic gradient descent based learning algorithm. (iii) We validate the model on four challenging datasets of expression recognition [24, 50], clinical pain prediction [25] and intent prediction (in dyadic conversations) [35]. We show that the method consistently outperforms temporal pooling and MIL based competitive baselines. In combination with complementary features, we report state-of-the-art results on these datasets with the proposed model.

2. Related works

Early approaches for facial expression recognition used apex (maximum expression) frames [38, 29, 5] or pre-segmented clips, and thus were strongly supervised. Also, they were often evaluated on posed video datasets [24].

To encode the faces into numerical vectors, many successful features were proposed e.g. Gabor [19] and Local Binary Patterns (LBP) [29], fiducial points based descriptors [49]. They handled videos by either aggregating features over all frames, using average or max-pooling [15, 36], or extending features to be spatio-temporal e.g. 3D Gabor [46] and LBP-TOP [51]. Facial Action Units, represent movement of facial muscle(s) [5], were automatically detected and used as high level features for video prediction [5, 20].

Noting that temporal dynamics are important for expressions [5], the recent focus has been more on algorithms capturing dynamics e.g. Hidden Markov Model (HMM) [4, 18] and Hidden Conditional Random Fields (HCRF) [3, 27, 31] have been used for predicting expressions. Chang et al. [3] proposed a HCRF based model that included a partially observed hidden state at the apex frame, to learn a more interpretable model where hidden states had specific meaning. The models based on HCRF are also similar to latent structural SVMs [43, 39], where the structure is defined as a linear chain over the frames. Other discriminative methods were proposed based on Dynamic Bayesian Networks [48] or hybrids of HMM and SVM [40]. Lorincz et al. [22] explored time-series kernels e.g. based on Dynamic Time Warping (DTW) for comparing expressions. Another model used probabalistic kernels for classifying exemplar HMM models [36].

Nguyen et al. [28] proposed a latent SVM based algorithm for classifying and localizing events in a time-series. They later proposed a fully supervised structured SVM for predicting Action Unit segments in video sequences [39]. Our algorithm differs from [28], while they use simple MIL, we detect multiple prototypical segments and further learn their temporal ordering. MIL based algorithm has also been used for predicting pain [37]. In recent works, MIL has been used with HMM [45] and also to learn embedding for multiple concepts [33] for predicting facial expressions. Rudovic et al. [32] proposed a CRF based model that accounted for ordinal relationships between expression intensities. Our work differs from this work in handling weakly labeled data and modeling the ordinal sequence between sub-events (see §1).

We also note the excellent performances reached by recurrent neural networks on video classification tasks e.g. Karpathy et al. [12] and the reference within. While such, neural networks based, methods lead to impressive results, they require a large amount of data to train. In the tasks we are interested in, collecting large amounts of data is costly and has practical and ethical challenges e.g. clinical pain prediction [25, 44]. While networks trained on large datasets for identity verification have been recently made public [30], we found empirically that they do not generalize effectively to the tasks we are interested in (§4).

3. Approach

We now describe our proposed Latent Ordinal Model (LOMo) in detail. We denote the video as a sequence of N frames[1] represented as a matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ with $\mathbf{x}_f \in \mathbb{R}^d$ being the feature vector for frame f. We work in a weakly supervised binary classification setting, where we are given a training set

$$\mathcal{X} = \{(X, y)\} \subset \mathbb{R}^{d \times N} \times \{-1, +1\} \quad (1)$$

containing videos annotated with the presence ($y = +1$) or absence ($y = -1$) of a class in X, without any annotations for specific columns of X i.e. $\mathbf{x}_f \forall f \in [1, N]$. While we present our model for the case of face videos annotated with absence or presence of an expression, we note that it is a general multi-dimensional vector sequence classification model.

---

[1] We assume, for brevity, all videos have the same number of frames, extension to different number of frames is immediate

---

Algorithm 1 SGD based learning for LOMo
___
1: *Given*: $\mathcal{X}, M, \lambda, \eta, k$
2: *Initalize*: $\mathbf{w}_i \leftarrow 0.01 \times \texttt{rand}(0, 1) \forall i \in [1, M], \mathbf{c} \leftarrow \mathbf{0}$
3: **for all** $t = 1, \ldots, \texttt{maxiter}$ **do**
4:     Randomly sample $(X, y) \in \mathcal{X}$
5:     Obtain $s_\Theta(X)$ and $\mathbf{k}$ using Eq. 4a
6:     **if** $y s_\Theta(X) < 1$ **then**
7:         **for all** $i = 1, \ldots, M$ **do**
8:             $\mathbf{w}_i \leftarrow \mathbf{w}_i(1 - \lambda\eta) + \frac{1}{M}\eta y_i \mathbf{x}_{k_i}$
9:         **end for**
10:         $c_{\sigma(\mathbf{k})} \leftarrow c_{\sigma(\mathbf{k})} - \eta$
11:     **end if**
12: **end for**
13: *Return*: Model $\Theta = \left(\{\mathbf{w}_i\}_{i=1}^M, \{c_j\}_{j=1}^{M!} \right)$
___

The model is a collection of discriminative templates (cf. SVM hyperplane parameters) and a cost function associated with the sequence of templates. The templates capture the appearances of different sub-events e.g. neutral, onset or offset phase of an expression [39], while the cost function captures the likelihood of the occurrence of the sub-events in different temporal orders. The parts and the cost function are all automatically and jointly learned, from the training data. Hence, the sub-events are not constrained to be either similar or distinct and are not fixed to represent certain expected states. They are mined from the data and could potentially be a combination of the sub-events generally used to describe expressions.

Formally, the model is given by

$$\Theta = \left(\{\mathbf{w}_i\}_{i=1}^M, \{c_j\}_{j=1}^{M!} \right), \mathbf{w}_i \in \mathbb{R}^d, c_j \in \mathbb{R} \quad (2)$$

with $i = 1, \ldots, M$ indexing over the M sub-event templates and $j = 1, \ldots, M!$ indexing over the different temporal orders in which these templates can occur. The cost function depends only on the ordering in which the sub-events occur in the current video, and hence is a look-up table (simple array, $\mathbf{c} = [c_1, \ldots, c_{M!}]$) with size equal to the number of permutations of the number of sub-events M. The reason and use of this will become more clear in §3.1 when we describe the scoring function.

We learn the model Θ with a regularized max-

margin hinge loss minimization, given by

$$\frac{\lambda}{2}\sum_{i=1}^{M}\|\mathbf{w}_i\|^2 + \frac{1}{|\mathcal{X}|}\sum_{X\in\mathcal{X}}[1 - y_i s_\Theta(X)]_+ \quad (3)$$

where $[a]_+ = \max(a, 0)\ \forall a \in \mathbb{R}$. $s_\Theta(X)$ is our scoring function which uses the templates and the cost function to assign a confidence score to the example X. The decision boundary is given by $s_\Theta(X) = 0$.

3.1. Scoring function

Deviating from a linear SVM classifier, which has a single parameter vector, our model has multiple such vectors which act at different temporal positions. We propose to score a video X, with model Θ, as

$$s_\Theta(X) = \max_{\mathbf{k}} \frac{1}{M}\sum_{i=1}^{M}\mathbf{w}_i^\top \mathbf{x}_{k_i} + c_{\sigma(\mathbf{k})} \quad (4a)$$

$$\text{s.t. } \mathcal{O}(\mathbf{k}) \leq \beta \quad (4b)$$

where, $\mathbf{k} = [k_1, \ldots, k_M] \in \mathbb{N}^M$ are the M latent variables, and $\sigma : \mathbb{N}^M \rightarrow \mathbb{N}$ maps $\mathbf{k} = (k_1, \ldots, k_M)$ to an index, with lexicographical ordering e.g. with $M = 4$ and without loss of generality $k_1 < k_2 < k_3 < k_4$, $\sigma(k_1, k_2, k_3, k_4) = 1, \sigma(k_1, k_2, k_4, k_3) = 2, \sigma(k_1, k_3, k_2, k_4) = 3$ and so on. The latent variables take the values of the frames on which the corresponding sub-event templates in the model gives maximal response while being penalized by the cost function for the sequence of occurrence of the sub-events. $\mathcal{O}(\mathbf{k})$ is an overlap function, with β being a threshold, to ensure that multiple $\mathbf{w}_i$'s do not select close by frames.

Intuitively, we capture the idea that each expression or pain sequence is composed of a small number of prototypical appearances e.g. onset and offset phase for smile, brow lower and cheek raise for pain, or a combination thereof. Each of the $\mathbf{w}_i$ captures such a prototypical appearance, albeit (i) they are learned in a discriminative framework and (ii) are mined automatically, again with a discriminative objective. The cost component c effectively learns the order in which such appearances should occur. It is expected to support the likely order of sub-events while penalizing the unlikely ones. Even if a negative example gives reasonable detections of such prototypical appearances, the order of such false positive detections is expected to be incorrect and it is expected to be penalized by the order dependent cost. We later validate such intuitions with qualitative results in §4.3.

3.2. Learning

We propose to learn the model using a stochastic gradient descent (SGD) based algorithm with analytically calculable sub-gradients. The algorithm, summarized in Alg. 1, randomly samples the training set and does stochastic updates based on the current example. Due to its stochastic nature, the algorithm is quite fast and is usable in online settings where the data is not entirely available in advance and arrives with time.

We solve the scoring optimization with an approximate algorithm. We obtain the best scoring frame $\mathbf{x}_{k_i}$ for $\mathbf{w}_i$ and remove $\mathbf{w}_i$ from the model and $\mathbf{x}_{f-t}, \ldots, \mathbf{x}_{f+t}$ frames from the video; and repeat steps M times so that every $\mathbf{w}_i$ has a corresponding $\mathbf{x}_{k_i}$. t is a hyperparameter to ensure temporal coverage by the model – it stops multiple $\mathbf{w}_i$'s from choosing (temporally) close frames. Once the $\mathbf{k} = k_1, \ldots, k_M$ are chosen we add $c_{\sigma(\mathbf{k})}$ to their average template score.

4. Experimental Results

We empirically evaluated the proposed approach on four challenging, publicly available, facial behavior datsets, of emotions, clinical pain and non-verbal behavior, in a weakly supervised setting i.e. without frame level annotations. The four datasets ranged from both posed (recorded in lab setting) to spontaneous expressions (recorded in realistic settings). We now briefly describe the datasets with experimental protocols used and the performance measures reported.

In the following, we first describe the datasets and their respective protocols and performance measures. We then give quantitative comparisons with out own implementation of competitive existing methods. We then present some qualitative results highlighting the choice of subevents and their orders by the method. Finally, we compare the proposed method with state-of-the-art methods on the datasets used.

CK+[2] **[24]** is a benchmark dataset for expression recognition, with 327 videos from 118 participants posing for seven basic emotions – anger, sadness, disgust, contempt, happy, surprise and fear. We use a standard subject independent 10 fold cross-validation and report mean of average class accuracies over the 10 folds. It has annotation for the apex frame and thus

---

[2] http://www.consortium.ri.cmu.edu/ckagree/

also allows fully supervised training and testing.

Oulu-CASIA VIS[3] **[50]** is another challenging benchmark for basic emotion classification. We used the subset of expressions that were recorded under the visible light condition. There are 480 sequences (from 80 subjects) and six classes (as CK+ except contempt). It has a higher variability due to differences among subjects. We report average accuracy across all classes and use subject independent folds provided by the dataset creators.

UNBC McMaster Shoulder Pain[4] **[25]** is used to evaluate clinical pain prediction. It consists of real world videos of subjects with pain while performing guided movements of their affected and unaffected arm in a clinical interview. The videos are rated for pain intensity (0 to 5) by trained experts. Following [45], we labeled videos as 'pain' for intensity above three and 'no pain' for intensity zero, and discarded the rest. This resulted in 149 videos from 25 subjects with 57 positive and 92 negative samples. Following [45] we do a standard leave-one-subject out cross-validation and report classification rate at ROC-EER.

LILiR[5] **[35]** is a dataset of non-verbal behavior such as agreeing, thinking, in natural social conversations. It contains 527 videos of 8 subjects involved in dyadic conversations. The videos are annotated for 4 displayed non-verbal behavior signals- agreeing, questioning, thinking and understanding, by multiple annotators. We generated positive and negative examples by thresholding the scores with a lower and higher value and discarding those in between. We then generated ten folds at random and report average Area under ROC – we will make our cross-validation folds public. This differs from Sheerman et al. [35], who used a very small subset of only 50 video samples that were annotated with the highest and the lowest scores.

4.1. Implementation Details and Baselines

We now give the details of the features used, followed by the details of the baselines and the parameter settings for the model learning algorithms (proposed and our implementations of the baselines).

Features. For our experiments, we computed four types of facial descriptors. We extracted 49 facial landmark points and head-pose information using supervised gradient descent[6] [47] and used them for aligning faces. The first set of descriptors were SIFT-based features, which we computed by extracting SIFT features around facial landmarks and thereafter concatenating them [47, 5]. We aligned the faces into 128×128 pixel and extracted SIFT features (using open source `vlfeat` library [41]) in a fixed window of size 12 pixels. The SIFT features were normalized to unit ℓ_2 norm. We chose location of 16 landmark points around eyes (4), brows (4), nose (2) and mouth (6) for extracting the features. Since SIFT features are known to contain redundant information [13], we used Principal Component Analysis to reduce their dimensionality to 24. To each of these frame-level features, we added coarse temporal information by appending the descriptors from next 5 consecutive frames, leading to a dimensionality of 1920. The second features that we used were geometric features [49, 5], that are known to contain shape or location information of permanent facial features (e.g. eyes, nose). We extracted them from each frame by subtracting x and y coordinates of the landmark points of that frame from the first frame (assumed to be neutral) of the video and concatenating them into a single vector (98 dimensions). We also computed LBP features[7] (with radius 1 and neighborhood 8) that represent texture information in an image as a histogram. We added spatial information to the LBP features by dividing the aligned faces into a 9×9 regular grid and concatenating the histograms (4779 dimensions) [38, 16]. We also considered Convolution Neural Network (CNN) features by using publicly available models of Parkhi et al. [30] that was trained on a large dataset for face recognition. We used the network output from the last fully connected layer. However, we found that these performed lower than other features e.g. on Oulu and CK+ datasets they performed about 10% absolute lower than LBP features. We suspected that they are not adapted to tasks other than identity discrimination and did not use them further.

Baselines. We report results with 4 baseline approaches. For first two baselines we used average (or mean) and max temporal pooling [36] over per-frame facial features along with SVM. Temporal pooling is often used along with spatio-temporal features such as Bag of Words [15, 37], LBP [51] in video event classification, as it yields vectorial representation for each video by summarizing variable length frame features. We selected Multiple Instance Learning based on latent SVM [2] as the third baseline algorithm. We also computed the performance of the fully supervised algorithms for cases with known location of the frame that contains the expression. For making a fair comparison, we used the same implementation for SVM, MIL and LOMo.

Parameters. We fix $M = 1$ and $c_\sigma = 0$ in the current implementation, for obtaining SVM baseline results with a single vector input, and report best results across both learning rate and number of iterations. For both MIL ($M = 1$) and LOMo, which take a sequence of vectors as input, we set the learning rate to $\eta = 0.05$ and for MIL we set $c_\sigma = 0$. We fix the regularization parameter $\lambda = 10^{-5}$ for all experiments. We do multiclass classification using one-vs-all strategy. For ensuring temporal coverage (see §3.2), we set the search space for finding the next sub-event to exclude $t = 5$ and 50 neighboring frames from the previously detected sub-events' locations for datasets with fewer frames per video (i.e. CK+, Oulu-CASIA VIS and LILiR datasets) and UNBC McMaster dataset, respectively. For our final implementation, we combined LOMo models learned on multiple features using late fusion i.e. we averaged the scores.

4.2. Quantitative Results

The performances of the proposed approach, along with those of the baseline methods, are shown in Table 1. In this comparison, we used SIFT-based facial features for all datasets. Since head nod information is important for identifying non-verbal behavior such as agreeing, we also appended head-pose information (yaw, pitch and roll) to the SIFT-based features for the LILiR dataset.

We see performance improvements with proposed LOMo, in comparison to baseline methods, on 6 out of 7 prediction tasks. In comparison to MIL, we ob-serve that LOMo outperforms the former method on all tasks. The improvements are 1.2%, 4.2% and 1.1% absolute, on CK+, Oulu-CASIA VIS and UNBC Mc-Master datasets, respectively. This improvement can be explained by the modeling advantages of LOMo, where it not only discovers multiple discriminative sub-events but also learns their ordinal arrangement. For the LILiR dataset, we see improvements in particular on the 'Questioning' (5.9% absolute) and 'Agreeing' (1.7% absolute), where temporal information is useful for recognition. In comparison to temporal pooling based approaches, LOMo outperforms both mean and max pooling on 6 out of 7 tasks. This is not surprising since temporal pooling operations are known to add noise to discriminative segments of a video by adding information from non-informative segments [36]. Moreover, they discard any temporal ordering, which is often important for analyzing facial activity [37].

On both facial expression tasks, i.e. emotion (CK+ and Oulu-CASIA VIS) and pain prediction (UNBC McMaster), methods can be arranged in increasing order of performance as mean-pooling, max-pooling, MIL, LOMo. A similar trend between temporal pooling and weakly supervised methods has also been reported by previous studies on video classification [37, 8]. We again stress that LOMo performs better than the existing weakly supervised methods, which are the preferred choice for these tasks. In particular, we observed the difference to be higher between temporal pooling and weakly supervised methods on the UNBC McMaster dataset, 67.4% for mean-pooling, 81.5% for max-pooling, 85.9% for MIL and 87.0% for LOMo. This is because the subjects exhibit both head movements and non-verbal behavior unrelated to pain, and thus focusing on the discriminative segment, cf. using a global description, leads to performance gain. However, we didn't notice a similar trend on the LILiR dataset – the differences are smaller or reversed e.g. for 'Understanding' mean-pooling is marginally better than MIL (79.4% vs. 78.9%), while LOMo is better than both (80.3%). This could be because most conversation videos are pre-segmented and predicting non-verbal behavior relying on a single prototypical segment might be difficult e.g. 'Understanding' includes both upward and downward head nod, which cannot be captured well by detecting a single event.

Dataset	Task	Full Sup.	Mean Pool	Max Pool	MIL	LOMo
Cohn-Kanade+	Emotion	91.9	86.0	87.5	90.8	**92.0**
Oulu-CASIA VIS	Emotion	75.0	68.3	69.0	69.8	**74.0**
UNBC McMaster	Pain	–	67.4	81.5	85.9	**87.0**
LILiR	Agree	–	84.7	**85.5**	77.7	79.4
	Question	–	86.2	84.3	80.7	**86.6**
	Thinking	–	93.6	88.9	93.8	**94.8**
	Understand	–	79.4	79.2	78.9	**80.3**

Table 1: Comparison of LOMo with Baseline methods on 4 facial behavior prediction datasets using SIFT based facial features (see §4.1).

In such cases we see LOMo beats MIL by temporal modeling of multiple events.
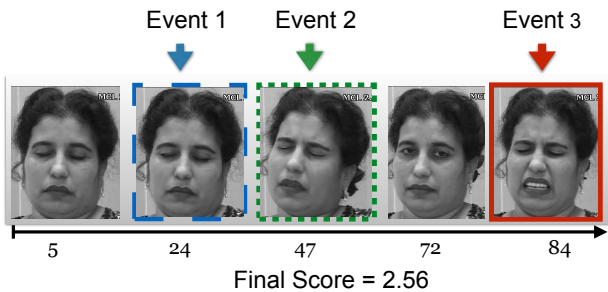

Figure 2: Detection of multiple discriminative sub-events, discovered by LOMo, on a video sequence from the UNBC McMaster Pain dataset. The number below the timeline shows the relative location (in percentile of total number of frames).

4.3. Qualitative Results

Fig. 3 shows the detections of our approach, with model trained for 'happy' expression, on two sequences from the Oulu-CASIA VIS dataset. The model was trained with three sub-events. As seen in Fig. 3, the three events seem to correspond to the expected semantic events i.e. neutral, low-intensity and apex, in that order, for the positive example (left), while for the negative example (right) the events are incorrectly detected and in the wrong order as well. Further, the final scores assigned to the negative example is -2.87 owing to low detection scores as well as penalization due to incorrect temporal order. The cost learned, by the model, for the ordering $(3, 1, 2)$ was -0.6 which is much lower than 0.9 for the correct order of $(1, 2, 3)$. This result highlights the modeling strength of LOMo, where it learns both multiple sub-events and a prior on their temporal order.

Fig. 2 shows detections on an example sequence from the UNBC McMaster dataset where subjects could show multiple expressions of pain [37, 33]. The results show that our approach is able to detect such multiple expressions of pain as sub-events.

Thus, we conclude that qualitatively our model supports our intuition, that not only the correct sub-events but their correct temporal order is critical for high performance in such tasks.

4.4. Comparison with State-of-the-Art

In this section we compare our approach with several existing approaches on the three facial expression datasets (CK+, Oulu-CASIA VIS and UNBC McMaster). Tab. 2 shows our results along with many competing methods on these datasets. To obtain the best performance from the model, we exploited the complementarity of different facial features by combining LOMo models learned on three facial descriptors – SIFT based, geometric and LBP (see §4.1). We used late fusion for combination by averaging the prediction scores from each model. With this setup, we achieve state-of-the-art results on the three datasets. We now discuss some representative works.

Several initial methods worked with pooling the spatio-temporal information in the videos e.g. (i) LBP-TOP [51] – Local Binary Patterns in three planes (XY and time), (ii) HOG3D [14] – spatio-temporal gradients, and (iii) 3D SIFT [34]. We report results from Liu et al. [21], who used a similar experimental protocol. These were initial works and we see that their performances are far from current method e.g. compared to 81.2% for the proposed LOMo, HOG3D obtains 70.6% and LBPTOP obtains 72.1% on the Oulu-CASIA VIS dataset.
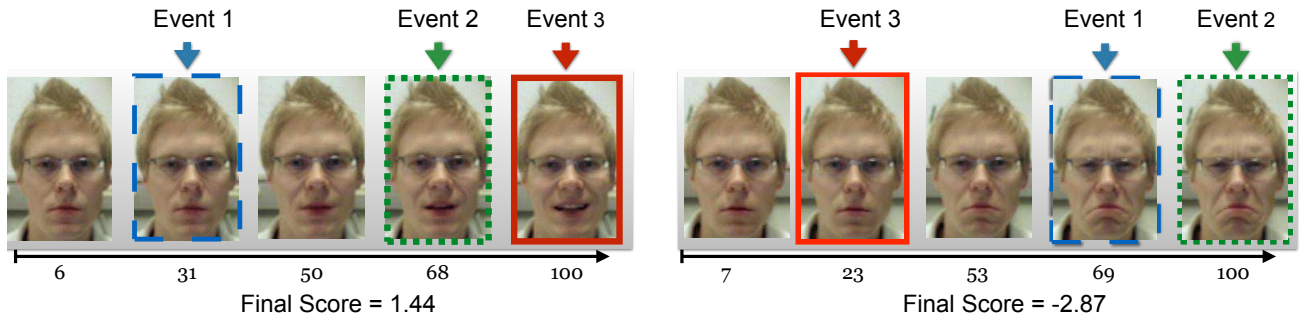
Figure 3: Detections made by LOMo trained ($M = 3$) for classifying 'happy' expression on two expression sequences from Oulu-CASIA VIS dataset. LOMo assigns a negative score to the sad expression (on the right) owing to negative detections for each sub-event and also negative cost of their ordering (see §3.1). The number below the timeline shows the relative location (in percentile of total number of frames).

CK+ dataset [24]		Oulu-CASIA VIS dataset [50]		UNBC McMaster dataset [25]	
3DSIFT [34]	81.4	HOG3D [14]	70.6	Ashraf et al. [26]	68.3
LBPTOP [51]	89.0	LBPTOP [51]	72.1	Lucey et al. [26]	81.0
HOG3D [14]	91.4	STM-ExpLet [21]	74.6	MS-MIL [37]	83.7
Ex-HMMs [36]	93.9	Atlases [9]	75.5	MIL-HMM [45]	85.2
STM-ExpLet [21]	94.2	Ex-HMMs [36]	75.6	RMC-MIL [33]	85.7
LOMo (proposed)	**95**.1	LOMo (proposed)	**82**.1	LOMo (proposed)	**87**.0

Table 2: Comparison of the proposed approach with several state-of-the-art algorithms on three datasets.

Approaches modeling temporal information include Exemplar-HMMs [36], STM-ExpLet [21], MS-MIL [42]. While Sikka et al. (Exemplar-HMM) [36] compute distances between exemplar HMM models, Liu et al. (STM-ExpLet) [21] learns a flexible spatio-temporal model by aligning local spatio-temporal features in an expression video with a universal Gaussian Mixture Model. LOMo outperforms such methods on both emotion classification tasks e.g. on Oulu-CASIA VIS dataset, LOMo achieves a performance improvement of 7.5% and 6.5% absolute relative to STM-ExpLet and Exemplar-HMMs respectively. Sikka et al. [37] first extracted multiple temporal segments and then used MIL based on boosting MIL [42]. Chongliang et al. [45] extended this approach to include temporal information by adapting HMM to MIL. We also note the performance in comparison to both MIL based approaches (MS-MIL [37] and MIL-HMM [45]) on the pain dataset. Both the methods report very competitive performances of 83.7% and 85.2% on UNBC McMaster dataset compared to 87.0% obtained by the proposed LOMo. Since having a large amount of data is difficult for many facial analysis tasks, e.g. clinical pain prediction, our results also show that combining, simple but complementary, features with a competitive model leads to higher results.

5. Conclusion

We proposed a (loosely) structured latent variable model that discovers prototypical and discriminative sub-events and learn a prior on the order in which they occur in the video. We learn the model with a regularized max-margin hinge loss minimization which we optimize with an efficient stochastic gradient descent based solver. We evaluated our model on four challenging datasets of expression recognition, clinical pain prediction and intent prediction is dyadic conversations. We provide experimental results that show that the proposed model consistently improves over other competitive baselines based on spatio-temporal pooling and Multiple Instance Learning. Further in combination with complementary features, the model achieves state-of-the-art results on the above datasets. We also showed qualitative results demonstrating the improved modeling capabilities of the proposed method. The model is a general ordered sequence prediction model and we hope to extend it to other sequence prediction tasks.

References

[1] F. Alnajar, Z. Lou, J. Alvarez, and T. Gevers. Expression-invariant age estimation. In *BMVC*, 2014.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[3] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *CVPR*, 2009.

[4] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *CVIU*, 91(1):160–187, 2003.

[5] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual analysis of humans*, pages 377–409. Springer, 2011.

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[8] A. Gaidon, Z. Harchaoui, and C. Schmid. Temporal Localization of Actions with Actoms. *PAMI*, 35(11):2782–2795, 2013.

[9] Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *ECCV*, 2012.

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[11] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. *Advances in Visual Computing*, pages 368–377, 2012.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[13] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *CVPR*, 2004.

[14] A. Klaser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[17] W. Li and N. Vasconcelos. Multiple instance learning for soft bags via top instances. In *CVPR*, 2015.

[18] J. J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Automated facial expression recognition based on facs action units. In *FG*, 1998.

[19] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *FG*, 2011.

[20] G. Littlewort, J. Whitehill, T.-F. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett. The motion in emotiona cert based approach to the fera emotion challenge. In *FG*, 2011.

[21] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, 2014.

[22] A. Lorincz, L. Jeni, Z. Szabo, J. F. Cohn, T. Kanade, et al. Emotional expression classification using time-series kernels. In *CVPRW*, 2013.

[23] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *PAMI*, 36(2):331–345, 2014.

[24] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[25] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*, 2011.

[26] P. Lucey, J. Howlett, J. Cohn, S. Lucey, S. Sridharan, and Z. Ambadar. Improving pain recognition through better utilisation of temporal information. In *AVSP*, 2008.

[27] D. McDuff, R. El Kaliouby, D. Demirdjian, and R. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, pages 1–7, 2013.

[28] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *CVPR*, 2009.

[29] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[31] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *PAMI*, 29(10):1848–1852, 2007.

[32] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *CVPR*, 2012.

[33] A. Ruiz, J. Van de Weijer, and X. Binefa. Regularized multi-concept mil for weakly-supervised facial behavior categorization. In *BMVC*, 2014.

[34] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM MM*, 2007.

[35] T. Sheerman-Chase, E.-J. Ong, and R. Bowden. Feature selection of facial displays for detection of non verbal communication in natural conversation. In *IC-CVW*, 2009.

[36] K. Sikka, A. Dhall, and M. Bartlett. Exemplar hidden markov models for classification of facial expressions in videos. In *CVPRW*, 2015.

[37] K. Sikka, A. Dhall, and M. S. Bartlett. Classification and weakly supervised pain localization using multiple segment representation. *IVC*, 32(10):659–670, 2014.

[38] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *ECCVW*, 2012.

[39] T. Simon, M. H. Nguyen, F. De La Torre, and J. F. Cohn. Action unit detection with segment-based svms. In *CVPR*, 2010.

[40] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE SMC*, 42(1):28–43, 2012.

[41] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[42] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. *NIPS*, 18:1417, 2006.

[43] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *CVPR*, 2009.

[44] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In *BMVC*, 2013.

[45] C. Wu, S. Wang, and Q. Ji. Multi-instance hidden markov model for facial expression recognition. In *FG*, 2015.

[46] T. Wu, M. S. Bartlett, and J. R. Movellan. Facial expression recognition using gabor motion energy filters. In *CVPRW*, 2010.

[47] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[48] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *PAMI*, 27(5):699–714, 2005.

[49] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *FG*, 1998.

[50] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *IVC*, 29(9):607–619, 2011.

[51] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *PAMI*, 29(6):915–928, 2007.

[52] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.