Cumulative Attributes for Pain Intensity Estimation

Joy O. Egede
University of Nottingham, China
joy.egede@nottingham.edu.cn

Michel Valstar
University of Nottingham, UK
michel.valstar@nottingham.ac.uk

ABSTRACT

Pain estimation from face video is a hard problem in automatic behaviour understanding. One major obstacle is the difficulty of collecting sufficient amounts of data, with balanced amounts of data for all pain intensity levels. To overcome this, we propose to adopt Cumulative Attributes, which assume that attributes for high pain levels with few examples are a superset of all attributes of lower pain levels. Experimental results show a consistent relative performance increase in the order of 20% regardless of features used. Our final system significantly outperforms the state of the art on the UNBC McMaster Shoulder Pain database by using cumulative attributes with Relevance Vector Regression on a combination of features, including appearance, geometric, and deep learned features.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**; *Kernel methods*; *Neural networks*;

KEYWORDS

Pain estimation; Attribute learning; Multi-output regression; Relevance Vector Machines (RVM)

ACM Reference format:
Joy O. Egede and Michel Valstar. 2017. Cumulative Attributes for Pain Intensity Estimation. In *Proceedings of ICMI '17, Glasgow, United Kingdom, November 13–17, 2017,* 8 pages.
https://doi.org/10.1145/3136755.3136789

1 INTRODUCTION

Pain assessment is an important part of clinical diagnosis. In clinical practice, a patient's pain level is routinely measured using standard pain assessment tools e.g. the Visual Analogue scale (VAS, [26]) and the Numeric Rating scale [39]. The VAS is mostly used for children and consists of describing faces ranging from 'no pain' to 'extreme pain'. Similarly, the numerical scale consists of a scale of numbers with the least and highest number signifying 'no pain' and 'maximum pain', respectively. Though useful, these tools rely on the patient being able to answer questions and are not applicable in cases where the patient is incapable of self-report e.g. newborns and unconscious patients. In such cases, pain assessment is done by proxy via observation of behavioural and physiological changes in the patient. However, the judgement made by the observer on the pain state of the patient can be influenced by a number of external factors e.g. training, experience, fatigue etc. Manual pain assessments also do not allow for continuous monitoring of the patient. This places enormous demands on the medical staff, especially for critically ill patients who need to be monitored more frequently.

Due to the limitations of the current clinical assessment process, machine learning techniques for automatic human behaviour understanding have been explored in order to introduce an objective, continuous-time and continuous-value pain estimate. This has mostly focused on the analysis of audio-visual recordings and physiological signals, both individually and in combination. Automatic pain recognition started from binary pain recognition i.e. detecting pain or no pain and progressed to continuous value pain estimation which is more useful for clinical assessment. Even though considerable advances have been made in this field, the performance of pain recognition systems is still limited.

For a large part, this is due to the relatively small datasets available for training. Due to privacy and ethical concerns, pain data is particularly difficult to obtain and where this is available there is the additional problem of sparse representation of the higher pain levels. For example, the McMaster database which is popular in pain expression analysis exhibits a high data imbalance with only 17.29% of 'pain' frames and 87.21% of 'no pain' frames (See Fig. 2). More so, only 0.85% of the data corresponds to pain levels 7 and higher. This imbalance negatively impacts on the ability of pain recognition systems to predict high pain levels accurately. Consequently, there is a need to harness the information from the abundant lower pain examples to generate useful data characteristics that can be exploited to make more accurate predictions of the higher pain levels, which have limited or no sample representation.

To address the data sparsity problem, we introduce Cumulative Attributes (CA) for Pain Estimation. CA models were originally developed to tackle the problem of data sparsity in age estimation by leveraging the shared characteristics between different classes. Cumulative Attributes have been used in age and crowd density estimation [6] but to the best of our knowledge, no study has considered its application to pain estimation or facial expression recognition in general. We propose here that automatic pain estimation is similarly an excellent target domain for Cumulative Attributes, and show this through a system attaining state of the art performance.

As in ageing, high levels of pain usually include all the signs of lower levels of pain, plus some extra features. This is particularly evident if one uses the well-known PSPI score [30], which is a summation of the intensities of the facial action units (FACS AUs, [9]) associated with pain i.e. the pain score increases as more AU elements are added or activated. Thus we formulate binary attributes corresponding to each pain level such that the attributes are both discriminative and capture the similarities between neighbouring pain levels. Specifically, for our 16-point PSPI scale, we formulate a

ICMI '17, November 13–17, 2017, Glasgow, United Kingdom

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136789

15-dimensional attribute vector such that for a pain level of 3, all the bits from 1 to the 3rd bit are set to 1 while all others are set to 0. For the special case of pain level zero, all the binary attributes are set to zero. Similar to [6], after generating the Cumulative Attributes from the target pain levels of the training set, we perform pain estimation using a two-level regression approach. First, we map the low-level appearance and shape features to the intermediate attribute space using multi output ridge regression. In the second stage, we map the learned attribute representation to the discrete pain values using a single output regression model. An advantage of this approach is that it implicitly performs feature dimensionality reduction i.e. it maps the high dimensional low-level features to a much lower dimensional attribute space.

Significant performance gains in facial expression analysis have been reported by the fusion of shape and appearance features, and recently it has been shown that adding Deep Learned features to hand-crafted features outperforms systems with only Deep-Learned features, in situations where the amount of training data is relatively small [8]. Thus we also implement a late fusion of the hand-crafted shape and appearance CA features with Deep Learned CA features. Pain expression involves multi-modal cues e.g. facial expressions, body movements, audible cues and changes in physiological signals. The proposed CA framework can readily be extended to include cues from multi-modalities as indicated by the shaded portion in Figure 1. We evaluate the proposed CA features on the UNBC McMaster database and show that CA performs significantly better than the base features. In addition, we compare with the state of the art and again show a significant improvement.

In summary, the contribution of this paper are threefold:

- We propose a Cumulative Attribute model for pain estimation where the attributes both discriminate and capture the common characteristics among neighbouring classes.
- We show that it consistently outperforms models trained on low level features.
- We attain results improving on the state of the art.

The rest of the article is structured as follows: Section 2 discusses the previous work on automatic pain recognition and also provides an overview of the concept of attribute learning in relation to object description and recognition. Section 3 presents our proposed Cumulative Attribute learning for pain estimation. It describes the process flow from low-level feature extraction to Cumulative Attribute learning and then pain estimation. In section 4, we describe the evaluation database in terms of its sparsity and data imbalance. We show the results on the learned Cumulative Attribute features in comparison to the low-level features and also provide a comparison with state-of-the-art methods.

2 RELATED WORK

The method proposed in this work merges the concept of attribute learning with automatic pain recognition. Thus, this section discusses the evolution of visual attributes with respect to object recognition and then extends to historical and current achievements on automatic pain recognition.

2.1 Automatic pain recognition

Automatic pain recognition has received growing attention in the past few years. Pain recognition is generally based on the analysis of pain indicators such as cry characteristics [5, 16], facial expressions [2, 4, 21, 23], body posture, [3] physiological signals [13, 28] or in combination [38]. Depending on the pain indicators and pain metric employed, previous work on automatic pain recognition can be classified into two categories: PSPI based and non-PSPI based pain recognition. We now discuss related work on pain estimation based on this classification.

PSPI based Pain recognition: Pain recognition from facial expressions is mostly based on the Prkachin and Solomon Pain Intensity (PSPI)[30] metric which measures pain as a linear combination of facial action units as shown in Eq. (1):

$$\begin{aligned} PSPI =& AU4 + max(AU6; AU7) + \\ & max(AU9; AU10) + AU43 \end{aligned} \tag{1}$$

Facial action units in this context refer to the contraction or expansion of specific facial muscles involved in a facial expression. Pioneering work [2, 23] on face based pain recognition assumed a binary classification problem i.e. pain and no pain. Attempts have also been made to recognize posed pain expressions from real pain [4, 21]. Attention quickly shifted to continuous pain estimation as this was more useful in clinical practice.

The problem of continuous pain estimation has been addressed in [17, 18, 24, 29, 32, 42]. Kaltwang et al. proposed a three step approach to pain estimation. First, frame-wise low-level features of shape and appearance are extracted. Then Relevance vector machines are trained on each feature type for pain estimation which is then followed by a late fusion of the shape and appearance features. Zafar and Khan [42] learn a KNN classifier on geometric features extracted from 22 facial points. However, their method requires a prior annotation of the neutral face for each subject. A supervised descent method (SDM) combined with SIFT features is proposed in [29] for both continuous and binary classification.

Facial expression recognition models have been shown to perform better when temporal information is included in the learning process. Thus, Kaltwang et al.[18] used temporal features in combination with part based feature extraction. The face is first divided into a uniform grid of cells, from which Local binary patterns are extracted over a specified time window. Florea et al. [12] proposed a histogram of topographical features for pain estimation in a transfer learning framework. Exploiting the temporal progression of pain expression from neutral through the apex and then back to neutral, Zhong et al. [45] propose ordinal information for pain estimation.

Deep learning has widely been used for various computer vision problem but only very recently has it been explored for pain estimation due to the limited pain data available for training. Similar to [18], a recurrent convolutional network is presented in [46] which uses a time windowed flattened 1D frame features as input to the network. Egede et al.[8] implemented a fusion of handcrafted and deep learned features with an additional person-specific normalisation of predictions. Temporal information is encoded within the deep learned features and they show that this complementary information improves the recognition model.

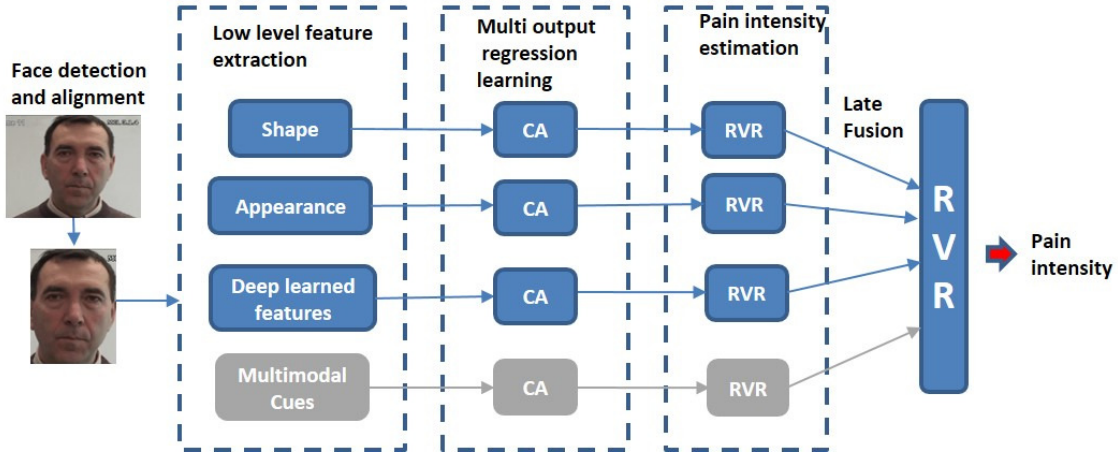

Figure 1: Block diagram of our proposed methodology. The grey blocks denote potential extension of the system with additional multimodal cues.

To deal with the problem of sparse data, previous works [33, 45] have resorted to clustering the pain levels for better performance i.e. the 16-point PSPI scale is reduced to a 5-point problem where levels 4-5 are merged while levels 6+ are grouped into one class. Doing this does not only simplify the task but also creates a more uniform dataset. The disadvantage of this approach is that the original goal of the PSPI is lost.

Non PSPI based Pain Recognition: Despite the achievements with video signals, there are still open challenges that limit its use in clinical settings. Recognition models can be adversely affected by factors such as poor illumination, out of plane head movements and face occlusion by medical devices. Consequently physiological [13, 28] and audio signals [5, 16] have been explored for pain estimation. Commonly used physiological signal includes respiration changes, galvanic skin response (GSR), electrocardiogram (ECG), photoplethysmogram and skin conductivity. An argument for this approach is that it is more reliable as these are directly controlled by the nervous system and cannot be easily faked unlike facial expressions. However, capturing these signals commonly involves invasive procedures where sensors have to be directly connected to the patient's body. Werner et al. [38] combine bio and visual signals for pain estimation. Evaluating on the BioVid database, they show that the combination performs better than the individual features.

2.2 Attribute learning

Attribute learning is a concept that has been extensively researched for object recognition. This is based on the notion that objects can be identified if the properties or attributes specific to that object are known. Visual attributes in this context are mid-level representations of objects such as shape, color, patterns, texture etc. However, manually annotating objects attribute can be time consuming and error prone. Secondly, annotations do not cover the true data space exhaustively as they only cover the available data. Problems arise when atypical examples are present in the test data, where 'atypical' is with respect to the training data. To deal with this problem, attributes that are learned from data were introduced.

A number of studies [10, 11, 36, 37] have focused on instance-level attribute prediction from data and shown that these attributes boost the performance of object classification models. Others [6, 20, 31, 41] have used class-level attributes in the sense that the attributes are crafted for each object class rather than for each training example. Class crafted attributes are much easier as in most cases classes are finite whereas instance-level attributes are more varied. Qin et al. [31] proposed a class-wise discrete descent algorithm that directly learns binary class level attributes which encode the between-class correlation and discriminative factors. In Yu et al. [41], a method is proposed for automatically constructing class-level attribute vectors where each element corresponds to a semantic label of the class.

One major advantage of learned attributes is their ability to generalise to unseen data. An example is the zero-shot learning concept which allows the identification of classes not present at training time. This is achieved in two steps: first, we learn the relationship between the attributes and the object features and then the learned attributes are used to map the objects to the target classes. This ability of data-learned attributes have been harnessed to address the problem of data sparsity and imbalance in challenging data sets where some of the classes have little or no data representation. Chen et al. [6] proposed a Cumulative Attribute model for age and crowd density estimation based on a two level regression learning. First, the attributes are learned simultaneously from the low-level features using a multi-output regression model. In the second level, the attributes are fed to single output regression model to predict the target scalar class i.e age or crowd count.

Most attribute learning techniques [10, 11, 20, 41] use a supervised learning approach where independent binary classifiers are learned for each attribute but then information on relationships between the attributes is lost. To preserve this information, models which implement a joint learning of the attributes have been proposed [1, 6, 27]. These supervised learning approaches require prior class or instance level attribute labels, but this might not always be available for real world data. To this end, Huang et al. [14] used an unsupervised deep learning framework to first identify 'classes' or

groups within an unlabelled data set. Then these learned classes are used as labels in a soft supervised approach to learn the shared attributes among the groups while maximizing the distance between the groups.

With the increased interest in deep learning for computer vision problems, deep neural networks have also been explored for generating attributes from data. A few studies [19, 22, 25, 35] have used Convolutional Neural Networks (CNNs) for facial attribute learning e.g. race, hair colour, eye glasses, smiling, facial make-up etc. In contrast to classifying faces based on learned attribute features, Sun et al. [35] first implement an identity recognition problem and then use the identity related features for attribute prediction. More recently, Zhang et al. [44] proposed a Tasks-Contrained Deep Convolutional Network (TCDCN) that optimizes facial landmark detection using shared representations learned from auxiliary attributes and show that this method performs better than CNN models learned directly from the image features.

To date, attribute learning has not been explored for pain expression analysis or facial expression recognition in general. Research in pain expression analysis is limited due to the difficulty associated with obtaining pain data. There is also the additional problem of data sparsity and high imbalance in the publicly available data sets. For example, in the McMaster shoulder pain database, which is annotated on the 16-point scale, only 0.05% of this belong to pain levels 7 and above. Due to the high imbalance and small data representation for higher pain levels, recognition models are unable to accurately predict high pain levels.

To deal with this problem of imbalance and data sparsity, we propose a Cumulative Attribute model for pain estimation from facial expressions. Cumulative Attributes were originally proposed for age and pain estimation due to the shared properties between the target classes, e.g. age. We argue that similar to age, neighbouring pain intensity levels share common characteristic which aggregate as one moves up the pain scale. This is particularly true for the PSPI which measures pain as a summation of facial action unit intensities. Thus, more Action Unit elements are added as the pain level increases. By leveraging on the shared characteristics between the pain levels, we can improve the model performance for pain levels which have low representation. An advantage of this approach is its simplicity in the sense that our attributes are class-based rather than instance-based. Hence, crafting the original attributes for learning is cost effective since our pain classes in this context are finite.

3 METHODOLOGY

In this section, we describe our Cumulative Attribute learning approach for pain estimation as shown in Figure 1. First, we construct Cumulative Attributes representations for each of the pain levels. Next, we extract low-level features from the face images and learn the relationship that maps these features to the attribute space using a multi-output regression framework. Using the learned multi-output regression function, we generate a new set of attribute features for all instances. Finally, we use this new attribute feature set to learn a linear regression function for continuous pain estimation.

3.1 Cumulative attributes learning for pain estimation

Consider the problem of predicting for every moment in time t, a label $y_t \in [0, 1, ..., k]$ and a set of features x_t to be extracted from the corresponding image. We construct a $k-1$ dimensional mid level attribute a_t such that:

$$a_t^j = \begin{cases} 1, & \text{when } \ j \leq y_t \\ 0, & \text{when } \ j > y_t. \end{cases}$$

Where $j = 1, 2, ..., k$ and k is the highest pain level. This implies that the first y_t elements in a_t are all ones while the last $k - y_t$ elements are zeros. In our case, $k = 15$ which corresponds to the highest pain level.

The motivation behind CA is that for some ordinal regression problem R, most observations of low values r_l of T are a subset of most higher values $r_h > r_l$. In addition, for many of these problems the number of examples n_l for low values is much greater than for high values of R, i.e. $n_l >> n_h$. It is therefore useful to have separate feature sets describing incrementally larger values of R. Age estimation, for which CA was originally proposed, is a prime example of this, as many features of ageing are permanent and are only added to when someone grows older (think of permanent wrinkles or the growing ears of men). Similarly, facial expressions of pain based on PSPI can be said to have a cumulative property, i.e. an expression of pain labelled as pain level k can be said to have all of the properties of pain level k-1 and below. Evidently, our Cumulative Attributes capture this incremental feature relationship while maintaining the distinction between the classes.

Following the procedure described in [6], we perform attribute learning in two stages. First, we learn the relationship that maps the low level features x to the attribute space a. To preserve the conditional dependence among the attributes, we jointly learn the attribute features using a multi-output regression model with a quadratic loss function. We find the parameters W and b that minimize the objective function:

$$arg_{w,b} min(\frac{1}{2}||W||_F^2 + C\sum_{i=1}^{N}||a_i^T - (x_i^T W + b)||_F^2$$

where $||.||_F^2$ is the frobenius norm, $\mathbf{a_i} = [a_i^1, a_i^2, ..., a_i^k]$ is the training attribute vector, $\mathbf{b} = [b^1, b^2, ..., b^k]$ is the bias term, $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^j, ..., \mathbf{w}^k]$ is the weight matrix and N is the total number of samples in the training set.

$\mathbf{W}$ and $\mathbf{b}$ are computed with a closed form solution as follows:

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{b} \end{bmatrix} = -(Q^T Q)^{-1} Q^T,$$

where Q and P are semi-definite matrices estimated as:

$$Q = \begin{bmatrix} 2C\sum_{i=1}^{N} x_i x_i^T & 2C\sum_{i=1}^{N} x_i \\ 2C\sum_{i=1}^{N} x_i^T & 2CN \end{bmatrix}$$

$$P = \begin{bmatrix} -2C\sum_{i=1}^{N} x_i a_i^T \\ -2C\sum_{i=1}^{N} a_i^T \end{bmatrix}$$

We determine the trade off parameter C by an inner loop cross validation on the training set. Finally, we learn the relationship

between the Cumulative Attribute space a and the output y using a single output regression function.

For this work, we have used two-modalities (i.e. changes in face shape and appearance) to estimate pain intensity. However, where multiple pain cues are available e.g. limb movement and audio, our CA architecture can be extended to accommodate new modalities. As shown in Figure 1, this can be achieved by extracting low-level features from the new modality and then learning the cumulative attributes.

3.2 Pain intensity estimation

To evaluate the performance of our attribute features for continuous pain estimation, we use Relevance Vector Machines to learn a regression function. RVM is a bayesian sparse kernel introduced to overcome the limitation of Support Vector Machines (SVM). Unlike SVMs, RVMs are capable of outputting posterior probabilities rather than just hard class decisions and have much sparser models. Similar to SVM, they are useful for both classification and regression problems. RVMS have been used in previous works [17, 18] on continuous pain estimation and have been shown to perform well. In [8], it was also shown that RVMs have a much lower minimum descriptor length with up to 28% reduction in the the number of decision vectors required across all features used in the experiment. To further support this, experiments by [40] show that RVMs are less susceptible to over fitting when compared to SVM. This is a desirable characteristic as we want our pain recognition model to generalize well to real world data.

To evaluate our method, we first learn an RVM regressor on each of the low-level shape, appearance and deep learned features. Good facial expression analysis requires a combination of shape and appearance features. Consequently, we do a late RVR feature fusion of the appearance and shape features such that the output of the single-feature RVMs are used as input to a second level RVR as seen in [17]. The same two-stage learning approach is used for pain estimation on the cumulative attribute features.

3.3 Fusion of Shape, Appearance and Deep learned features

Appearance and shape features have served as building blocks for facial expression analysis. Different variants of these have been used extensively both individually and in combination. It has been established that combining both features yields improved performance as they both capture intrinsic facial characteristics which are complementary. Appearance features capture facial deformations such as texture, furrows and wrinkles. Shape features, on the other hand, are built from the geometry of facial components such as the eyes, nose, mouth. In this work we use Histogram of Oriented Gradient (HOG) [7] descriptors in combination with distance metrics computed from the location and displacement of facial components in reaction to pain.

First, we obtain 66 facial points for each face image using the publicly available iCCR landmark tracker [34]. This is in contrast to previous work [8, 12, 17, 18, 33] which use the manually annotated facial points provided with the database. Facial points obtained from the tracker are used for face registration. The registered face images are then aligned to a mean shape based on a Procrustes transformation. Next, we extract features from the aligned face images. To extract the HOG features, we isolate a 24x24 pixel patch around each facial point. Each patch is further split into a 2x2 cell. 9 bins of oriented gradients are extracted from each cell resulting in 2376D feature representation for the input face image. Similar to [8], the shape features comprise of the aligned facial points and a number of distance metrics computed from the facial points. A detailed description of the distance metrics can be found in [8].

In addition to the hand-crafted features, we experimented with deep learned features. Since deep learned models do not work well with limited data as in our case, we use pre-trained Convolutional Neural Networks (CNN) for feature extraction. CNNs are feed-forward artificial neural networks which have been widely used in many machine learning applications e.g. video analysis, object detection, gaming and natural language processing. CNNs are not entirely a new concept but its recent popularity stemmed from the huge success recorded with CNNs in classification problems over hand-crafted features. CNNs are able to self-learn features or data characteristics relevant to the machine learning problem, thus eliminating the burden of coding these features manually.

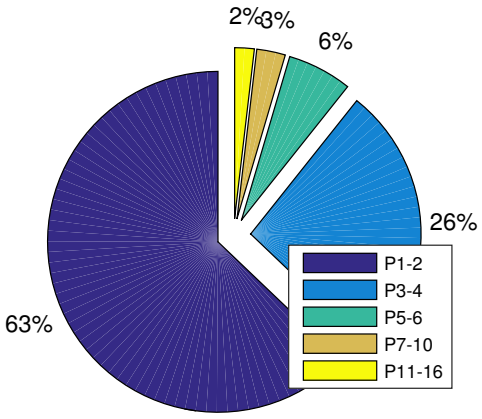

Figure 2: Percentage Distribution of non-zero pain levels in the McMaster database

Specifically, we use the CNN architecture proposed in [15] for Action Unit (AU) detection. The CNNs are pre-trained with images from the BP4D database [43]. A rationale for this is that the PSPI metric is a combination of action units hence it follows that CNNs trained for AU detection will produce features relevant for pain recognition. Two CNNs are separately trained for AUs corresponding to the eye and mouth region respectively. Details of the CNN input, architecture and training can be found in [15]. We extract deep learned features from the last 3072D fully connected convolution layers. Merging the features from both CNNs results in a 6144D representation for the face. In addition, the CNN features encode temporal information which complements our Cumulative Attributes. The shape, appearance and deep learned features are used for pain estimation both separately and in combination via a late feature fusion approach.

4 EXPERIMENTAL EVALUATION

4.1 Evaluation database

To enable comparison with previous work on continuous pain estimation, we evaluate our Cumulative Attributes on the publicly available McMaster shoulder Database. It consists of 200 video sequences of 25 subjects performing various arm movements following the instruction of a physiotherapist. The arm movements are carried out in two variations; the active mode where the patient move their arms and the passive mode, where the physiotherapist moves the arms within bearable pain limits. All video frames are fully annotated for Action Unit intensities with respect to the PSPI metric. The database also includes sequence level pain annotations on a 5-point scale making it an excellent setup for evaluating pain recognition techniques. However, only a small number of these frames contain pain expressions thus making it a challenging dataset to work with. Figure 2 shows the PSPI distribution of frames in the database.

The UNBC McMaster database is an ideal target for our proposed approach due to the high data imbalance and data sparsity for high pain levels. Even though it appears to have a high number of frame counts, over 80% of these do not have any pain signals. In addition, a high proportion of the pain signals frames are concentrated on the lower levels with only 0.85% of these belonging to pain levels 7 and above. We show that our method is robust to the sparsity problem and capable of achieving high performance on a continuous scale.

4.2 Experiments and Results

We performed two sets of experiments to evaluate our proposed CA framework. First, we performed pain estimation using the low-level features and this serves as our baseline for comparison. In the second set, we used the learned CA features. To support comparison with previous work, we used the Pearson correlation coefficient (PCC) and Root Mean Square Error (RMSE) for performance evaluation. RMSE is computed as the difference between the predictions and the ground truth which in this case are the PSPI annotations for each frame. Similarly, PCC measures how well the predictions mirrors the temporal progression of pain expression in the video sequences. A leave-one-out subject independent cross validation approach is used in all our experiments. The radial basis function (RBF) is used for the RVM Kernel as this have been shown to perform well for pain estimation [8, 17]. The length-scale parameter of the RVM kernel is determined by an inner-loop cross validation on the training set. To reduce the imbalance in the training set, we randomly sample from the zero-pain frames with a ratio of 1:2 corresponding to the modal pain signal frames and zero-pain frames respectively.

Figure 3 shows a comparison of the RVR performance on our Cumulative Attributes (CA) in relation to the non-Cumulative Attributes (NCA). NCA in this context refers to the original low-level features extracted from the face image. It can be seen that for all features, the Cumulative Attribute features consistently improve the performance of the RVM. Specifically, we achieved an average of 14% increase in Pearson correlation and an 18% reduction in RMSE across all features. This is particularly more evident with the single feature RVMs which were directly trained on attribute features. A possible explanation for the relatively high reduction

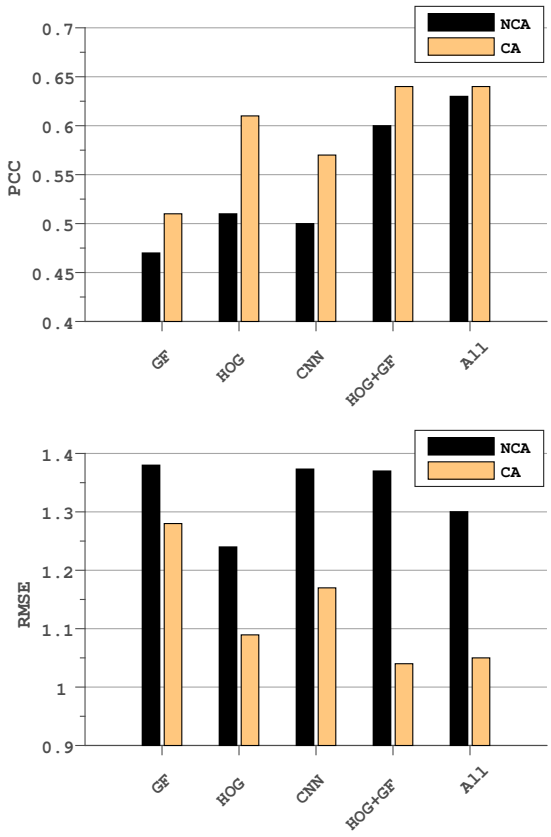

Figure 3: Comparison of RVM performance (PCC & RMSE) on the cumulative vs low level features

in RMSE is that with Cumulative Attributes, the model is able to predict relatively higher values for the high pain classes compared to the models based on low-level features.

In comparing the PCC for the CA and NCA fused-features, we observe that the performance improvement is not as much as that obtained with the single features. However, we observe a high improvement in the RMSE. A possible explanation for this is that the fused low-level features are able to capture the temporal progression of pain expression but unlike the CA features, it fails to effectively capture the magnitude of the pain. As such, the NCA fused features give higher error values in spite of the relatively good PCC. An

Table 1: Comparison of our Cumulative Attributes approach with the state-of-the-art

	RMSE	PCC
Kaltwang et al. [17]	1.18	0.59
Neshov & Manolava [29]	1.13	0.59
Kaltwang et al. [18]	1.69	0.66
Florea et al. [12]	1.10	0.53
Zhou et al. [46]	1.24	0.65
Egede et al. [8]	1.30	0.63
Our method	**1.04**	**0.64**

example of this occurrence can be seen in Figure 4 which shows the RVM prediction for both CA and NCA features in comparison to the ground truth on a time series scale. Evidently, the CA features perform better than the NCA features in predicting high pain levels and also has a better approximation for the lower pain levels.
To further confirm this, we show the average error rates for the pain levels in Figure 5. It can be seen that RMSE drops for the Cumulative Attribute features as we move higher up the pain levels and this effect is particularly evident for the HOG features.

In comparing the RVM performance on single versus fused features, we achieved a better performance with RVM fusion for both CA and non-CA experiments. This is in line with observations in [8, 12, 17] where such techniques have been shown to yield performance gains. Among the individual features, the HOG descriptor gives the best performance with the geometric features performing much worse.

Table 1 shows a comparison of our method with the state-of-the-art. We only include the performance from previous studies who have reported their results on a 16-point scale for a fair comparison. Some previous studies [33, 45] have adopted a simpler approach of reducing the original PSPI 16-point problem to a 6-point scale. The original 16 pain levels are discretized in a data balancing manner. Specifically, levels 4 & 5 are merged and pain levels 6 and above are categorized as one class. This is a much simpler task because most of the errors in prediction occur on the higher pain levels due to the small sample representation. It is not a surprise therefore, that they record very high performance figures. We do not include their results in the comparison table as the tasks are not comparable. Note that in comparing with [8], we report on their unprocessed results because our proposed approach does not include the additional person-specific normalization of the predictions which is implemented in [8]. Even though our PCC is slightly lower than

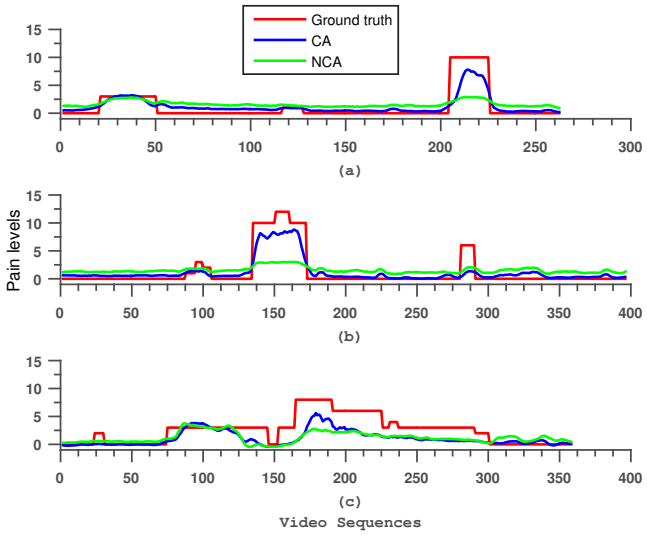

Figure 4: RVM predictions on CA vs low-level features on video sequences of Subjects with high pain occurrences (a) 064-ak064:ak064t1aaaff (b) 064-ak064:ak064t1aeunaff (c) 123-jh123:jh123t1aeaff

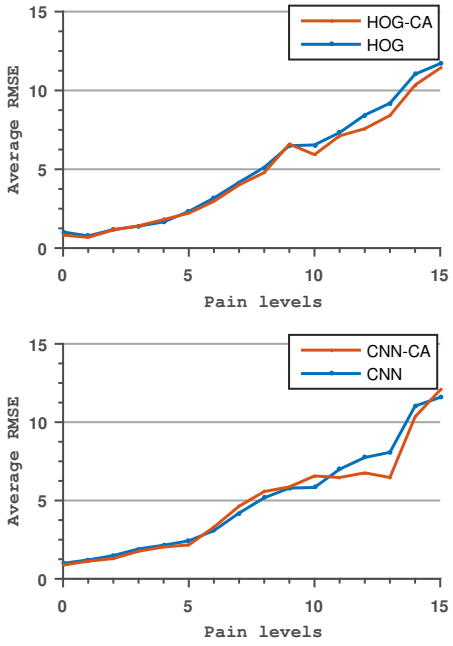

Figure 5: Comparison of average error rates on higher pain levels for CA vs low level features

[18] and [46] we achieve a 70% reduction in RMSE in comparison to [18] and 19% in comparison to [46] respectively. Note that, though a high PCC is desirable, it is also important to have a low prediction error rate. The two performance measures have to be considered together rather than individually as they are invariant i.e. a high PCC does not necessarily guarantee a lower error (RMSE) and vice versa. Our method not only performs better in terms of RMSE, it also gives a comparable PCC.

5 CONCLUSION

In this work, we introduced Cumulative Attributes for continuous pain intensity estimation. We learn Cumulative Attributes which encodes the shared characteristics between neighbouring pain levels and at the same time are distinct for each level. All attributes are learned simultaneously which preserves the conditional dependence among the attributes. We show that for all features types, the RVMs learned on the cumulative attributes performed consistently better in predicting higher pain levels when compared to those trained on the corresponding low-level features.

ACKNOWLEDGMENTS

The lead author acknowledges the financial support from the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, and the University of Nottingham. This work was also supported by the UK Engineering and Physical Sciences Research Council [grant number EP/L015463/1]. The work of Valstar is supported by European Union Horizon 2020 research and innovation programme under grant agreement No 645378.

REFERENCES

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. 2013. Label-Embedding for Attribute-Based Classification. In *2013 IEEE CVPR*. 819–826. https://doi.org/10.1109/CVPR.2013.111

[2] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M. Prkachin, and Patricia E. Solomon. 2009. The painful face – Pain expression recognition using active appearance models. *Image and Vision Computing* 27, 12 (nov 2009), 1788–1796. https://doi.org/10.1016/j.imavis.2009.05.007

[3] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. d. C. Williams, M. Pantic, and N. Bianchi-Berthouze. 2016. The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset. *IEEE TAC* 7, 4 (Oct 2016), 435–451. https://doi.org/10.1109/TAFFC.2015.2462830

[4] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, and Kang Lee. 2014. Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions. *Current Biology* 24, 7 (mar 2014), 738–743. https://doi.org/10.1016/j.cub.2014.02.009

[5] Chuan-Yu Chang and Jia-Jing Li. 2016. Application of deep learning for recognizing infant cries. In *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 1–2.

[6] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. 2013. Cumulative Attribute Space for Age and Crowd Density Estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE). https://doi.org/10.1109/cvpr.2013.319

[7] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE CVPR 2005.*, Vol. 1. IEEE, 886–893.

[8] J. Egede, M. Valstar, and B. Martinez. 2017. Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation. In *2017 12th IEEE International FG 2017*. 689–696. https://doi.org/10.1109/FG.2017.87

[9] P. Ekman, W.V. Friesen, and J.C. Hager. 2002. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City (USA).

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE CVPR*. 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772

[11] Vittorio Ferrari and Andrew Zisserman. 2007. Learning Visual Attributes.. In *NIPS*, Vol. 1. 2.

[12] Corneliu Florea, Laura Florea, and Constantin Vertan. 2015. *Learning Pain from Emotion: Transferred HoT Data Representation for Pain Intensity Estimation*. Springer International Publishing, Cham, 778–790. https://doi.org/10.1007/978-3-319-16199-0_54

[13] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. Springer.

[14] C. Huang, C. C. Loy, and X. Tang. 2016. Unsupervised Learning of Discriminative Attributes and Visual Representations. In *2016 IEEE CVPR*. 5175–5184. https://doi.org/10.1109/CVPR.2016.559

[15] Shashank Jaiswal and Michel Valstar. 2016. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE WACV*. IEEE, 1–8.

[16] Mahmoud Mansouri Jam and Hamed Sadjedi. 2009. A System for Detecting of Infants with Pain from Normal Infants Based on Multi-band Spectral Entropy by Infant's Cry Analysis. In *Computer and Electrical Engineering, 2009. ICCEE'09. Second International Conference on*, Vol. 2. IEEE, 72–76.

[17] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. 2012. Continuous Pain Intensity Estimation from Facial Expressions. In *Advances in Visual Computing*. Springer Science + Business Media, 368–377. https://doi.org/10.1007/978-3-642-33191-6_36

[18] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. 2015. Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation. (2015).

[19] S. Kang, D. Lee, and C. D. Yoo. 2015. Face attribute classification using attribute-aware correlation map and gated convolutional neural networks. In *2015 IEEE ICIP*. 4922–4926. https://doi.org/10.1109/ICIP.2015.7351743

[20] C. H. Lampert, H. Nickisch, and S. Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 951–958.

[21] Gwen C. Littlewort, Marian Stewart Bartlett, and Kang Lee. 2009. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing* 27, 12 (nov 2009), 1797–1803. https://doi.org/10.1016/j.imavis.2008.12.010

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE ICCV*. 3730–3738. https://doi.org/10.1109/ICCV.2015.425

[23] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. 2011. Automatically detecting pain in video through facial action units. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 41, 3 (2011), 664–674.

[24] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, Sien Chew, and Iain Matthews. 2012. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing* 30, 3 (2012), 197–205.

[25] P. Luo, X. Wang, and X. Tang. 2013. A Deep Sum-Product Architecture for Robust Facial Attributes Analysis. In *2013 IEEE ICCV*. 2864–2871. https://doi.org/10.1109/ICCV.2013.356

[26] M Lynch. 2001. Pain as the fifth vital sign. *Journal of intravenous nursing : the official publication of the Intravenous Nurses Society* 24, 2 (2001), 85âĂŤ94. http://europepmc.org/abstract/MED/11836838

[27] D. Mahajan, S. Sellamanickam, and V. Nair. 2011. A joint learning framework for attribute models and object descriptions. In *2011 International Conference on Computer Vision*. 1227–1234. https://doi.org/10.1109/ICCV.2011.6126373

[28] Fatma Nasoz, Kaye Alvarez, Christine L Lisetti, and Neal Finkelstein. 2004. Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work* 6, 1 (2004), 4–14.

[29] Nikolay Neshov and Agata Manolova. 2015. Pain detection from facial characteristics using supervised descent method. In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on*, Vol. 1. IEEE, 251–256.

[30] Kenneth M. Prkachin and Patricia E. Solomon. 2008. The structure reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain* 139, 2 (oct 2008), 267–274. https://doi.org/10.1016/j.pain.2008.04.010

[31] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao. 2016. Beyond Semantic Attributes: Discrete Latent Attributes Learning for Zero-Shot Recognition. *IEEE Signal Processing Letters* 23, 11 (Nov 2016), 1667–1671. https://doi.org/10.1109/LSP.2016.2612247

[32] Neeru Rathee and Dinesh Ganotra. 2015. A novel approach for pain intensity detection based on facial feature deformations. *Journal of Visual Communication and Image Representation* 33 (2015), 247–254.

[33] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. 2013. *Automatic Pain Intensity Estimation with Heteroscedastic Conditional Ordinal Random Fields*. Springer Berlin Heidelberg, Berlin, Heidelberg, 234–243. https://doi.org/10.1007/978-3-642-41939-3_23

[34] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. 2016. Cascaded Continuous Regression for Real-Time Incremental Face Tracking. In *ECCV 2016, Part VIII*. 645–661.

[35] Y. Sun, X. Wang, and X. Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *2015 IEEE CVPR*. 2892–2900. https://doi.org/10.1109/CVPR.2015.7298907

[36] Gang Wang and D. Forsyth. 2009. Joint learning of visual attributes, object classes and visual saliency. In *2009 IEEE 12th ICCV*. 537–544. https://doi.org/10.1109/ICCV.2009.5459194

[37] X. Wang, S. Qiu, K. Liu, and X. Tang. 2014. Web Image Re-Ranking Using Query-Specific Semantic Signatures. *IEEE Transactions on PAMI* 36, 4 (April 2014), 810–823. https://doi.org/10.1109/TPAMI.2013.214

[38] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C Traue. 2014. Automatic pain recognition from video and biomedical signals. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 4582–4587.

[39] Amanda C de C Williams, Huw Talfryn Oakley Davies, and Yasmin Chadury. 2000. Simple pain rating scales hide complex idiosyncratic meanings. *Pain* 85, 3 (2000), 457–463.

[40] Xu Xiang-min, Mao Yun-feng, Xiong Jia-ni, and Zhou Feng-le. 2007. Classification performance comparison between RVM and SVM. In *2007 International Workshop on Anti-Counterfeiting, Security and Identification (ASID)*. IEEE, 208–211.

[41] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S. F. Chang. 2013. Designing Category-Level Attributes for Discriminative Visual Recognition. In *2013 IEEE CVPR*. 771–778.

[42] Zuhair Zafar and Nadeem Ahmad Khan. 2014. Pain Intensity Evaluation through Facial Action Units. In *2014 22nd International Conference on Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE). https://doi.org/10.1109/icpr.2014.803

[43] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706.

[44] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. 2016. Learning Deep Representation for Face Alignment with Auxiliary Attributes. *IEEE Transactions on PAMI* 38, 5 (May 2016), 918–930. https://doi.org/10.1109/TPAMI.2015.2469286

[45] R. Zhao, Q. Gan, S. Wang, and Q. Ji. 2016. Facial Expression Intensity Estimation Using Ordinal Information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3466–3474. https://doi.org/10.1109/CVPR.2016.377

[46] J. Zhou, X. Hong, F. Su, and G. Zhao. 2016. Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1535–1543. https://doi.org/10.1109/CVPRW.2016.191