Beyond Deep Feature Averaging: Sampling Videos Towards Practical Facial Pain Recognition

Xiang Xiang *
Amazon Web Services, Inc.

Abstract

In hospitals, automatic identification of patients with cameras can greatly generalize the applicability of intelligent patient monitoring. However, patients unaware of being monitored do not adjust their behaviors, making pose variation a challenge. We argue that the frame-wise feature mean is unable to characterize the variation among frames. We propose to preserve the overall pose diversity if we want the video feature to represent the subject identity. Then identity will be the only source of variation across videos since pose varies even within a single video. Following that variation disentanglement idea, we present a pose-robust face verification algorithm with each video represented as an ensemble of frame-wise CNN features. Another challenge is that patients may move anytime, which makes real-time processing of a video stream a necessity. Instead of simply using all the frames, the algorithm is highlighted at the key frame selection by pose quantization using pose distances to K-means centroids, which reduces the number of feature vectors from hundreds to K while still preserving the overall diversity. We analyze how such a video sampling strategy is better than random sampling. An end-to-end face recognition algorithm is developed for real-time patient identification with a rank-list of one-to-one similarities using the proposed video representation. It works well in practice and generates a private patient dataset on the fly. On the official 5000 video-pairs of public YouTube Face dataset, our algorithm achieves a comparable performance with state-of-the-art that averages over deep features of all frames. In summary, the main contribution of this paper is a video-versus-video consensus with discriminative metric learning on the fly, which is verified in a working system for the patient monitoring system.

Figure 1: Painful expression can be subtle and short. Detection and measurement are difficult. Pain level is defined as AU4 + (AU6——AU7) + (AU9——AU10) + AU43 [18] [from the Prkachin and Solomon pain intensity (PSPI) metric].

1 Introduction

According to Wikipedia, patient originally meant 'one who suffers or has pain'. Obtaining accurate patient-reported pain assessments is especially important to effectively manage pain in the irradiated head neck cancer (HNC) patients. Previous work by our HNC clinic team has demonstrated a role for prophylactic pain management in HNC patients by requiring active pain monitoring to identify early increases in pain intensity to titrate the analgesic and achieve early pain control. This results in a reduction of the overall narcotic dose that is needed improving swallow function and is of particular interest given the national concern of opioid dependency, especially with increasing cancer survivors. However, facil-

*This work was done prior to Xiang joining Amazon AI.

itating self-reporting for patients at scale is difficult due to a lack of existing technological tools and the dynamic nature of patients self-assessments. Current solutions for this out-of-clinic pain assessment protocol lie in the development of a smartphone app, wherein patients enter pain levels. This limits adoptability (and accuracy) because it requires significant attention and effort by the patient in a scenario where they are under duress. To this end, this proposal seeks to develop an automated approach to pain assessment based on facial analysis from easily-obtainable video sequences. This serves two purposes: first, it simplifies the data collection process for the patient and reduce the strain on their manual efforts; second, it standardizes the feedback mechanism by ensuring one system perform all assessments and reduce bias, thereby enabling earlier intervention by clinicians to manage the pain for HNC patients.

We need the sample mean and variance to approximate a true data distribution while the sample mean itself is not a robust statistic. However, feature averaging is straightforward and conventional to represent a sequence such as in the recent video-based recognition works such as face recognition, activity recognition and video captioning. Taking hospitals as an instance, automatic identification of patients with a camera can greatly generalize the applicability of intelligent patient monitoring. However, patients unaware of being monitored do not adjust their behaviors, making pose variation a primary challenge. We argue that the frame-wise feature mean is unable to characterize the variation existing among frames. If we want the video feature to represent the subject identity, we had better preserve the overall pose diversity, because Convolutional Neural Networks (CNN) features are normally not robust to poses. Then, disregarding factors other than identity and pose, identity will be the only source of variation across videos since pose varies even within a single video. Following such a variation disentanglement idea, we present a pose-robust face recognition algorithm with each video represented as an ensemble of frame-wise CNN features. Moreover, patients may move anytime, making real-time processing of a video stream a necessity. Instead of simply using all the frames, the algorithm is highlighted at the key frame selection by pose quantization using pose distances to K-means centroids, which reduces the number of feature vectors from hundreds to K while still preserving the overall diversity. In particular, we analyze how such a video sampling strategy is better than random sampling. In this paper, an end-to-end face recognition system is presented for real-time patient identification with the proposed video representation. The correlation, a simple metric though, between two video features is employed to measure how likely the two videos represent the same person. The system runs well real-time and generates a private patient dataset on the fly. On the official 5000 video-pairs of public YouTube Face dataset, our algorithm achieves a comparable performance with state-of-the-art that averages over deep features of all frames.

2 Related Works

Face verification is a key subproblem of face recognition, e.g., it determines whether a pair of two face images are from the same person or not. A practical face verification system needs to run at real time with a high accuracy. Recently, CNN has shown significant improvements over traditional approaches in terms of the verification accuracy [1][2][10]. However, the computational overload of CNN makes it hardly usable for the practical video-based face verification systems. The main challenges of face verification arise from the high within-identity variations and the high identity-identity similarities.

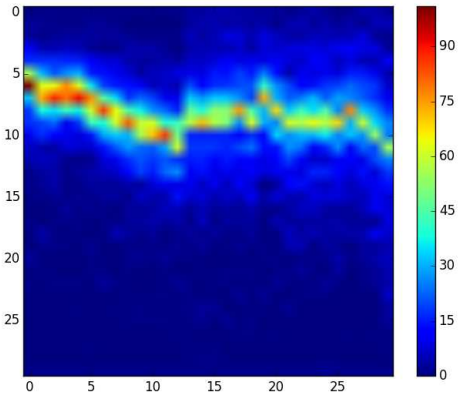

Figure 2: The histogram map of ECG pain signals over time. The X axis is the time. The Y axis is the bin number, from the smaller signal value in the top to the larger in the bottom.

There are many different types of face verifications. For Web-based applications, verification is conducted by comparing images to images. The images may be of the same person but were taken at different time or under different conditions. For online face verification, alive video rather than still images is used. More specifically, the existing video-based verification solutions assume that reference face images are taken under controlled conditions [19]. However, in practical scenarios, references are often taken uncontrolled, e.g. automatically taking photos of a customer with a camera deployed at the reception place of a hospitals or a hotel.

The conventional way of using handcrafted features such as Local Binary Patterns (LBP) [18] or Wavelet [] does not suffer from the low-speed issue. In this deep learning era [9], face recognition on a number of benchmarks such as LFW [4] has been well solved by DeepFace [2], DeepID [16], FaceNet [1] and so on. See VGGFace [10] and reference therein for a systematic review.

Usually we measure the similarity of the two subject and then make a decision by thresholding the similarity. The feature describing a subjects is extracted using either a shallow model [11][12] or a deep model [1][2][10]. We use a deep model in this paper. The cosine similarity or correlation both are well-defined similarity metrics. In contrast, we could also measure how different two subjects are. For example, the dissimilarity can be measured by Euclidean distance as it is anti-correlated with correlation.

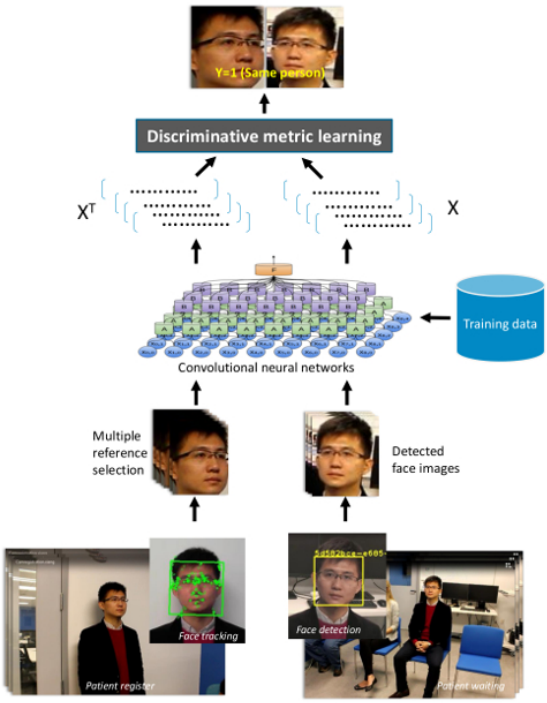

Figure 3: At registration (bottom-left), a inconspicuous webcam tracks the face of the patient and selects several face images as references for the patient. At testing (bottom-right), ceiling-mounted cameras continuously scan the crowd, detect faces and compare them with the reference images of all patients. The CNN feature of a image without face frontalization is hoped to be invariant to poses.

3 The Algorithm

3.1 Key face selection by pose quantization

Using a full live video stream will require many computational resources. Instead, we select up to n face keyframes out of a video stream according 3D poses. Intuitively, we want to retain the key faces that are as different as possible. Practically, we select the face keyframes under as-different-as-possible 3-D poses. We sample n reference images at registration and compute feature vectors $a_1, ..., a_n \in R^d$. At testing, for each frame we compute a feature vector $y \in R^d$. Since $a_1, ..., a_n$ are different representation of the same subject, they are correlated. We would like to maximize the diversity of reference image set. Thus, the objective of reference image selection is to

$$\max_{w} \sum_{i \neq j} (a_i - a_j)^T w w^T (a_i - a_j) \tag{1}$$

which is unsupervised learning of a subspace w from the data matrix A.

First, we compute the rotation angles of roll, pitch and yaw for each face video using existing 3D pose estimation method in OpenCV. Then, we performs a frame-wise vector quantization which reduces the number of images required to represent the face from tens or hundreds to K (say, K = 9 for a K-means codebook), while preserving the overall diversity.

3.2 CNN-based Representation

The pretrained VGG face model [] is used for our verification purpose without any re-training. The model has 24 layers, including several stacked convolution-pooling layer, 2 fully-connected layer and one softmax layer. Since the model was trained for classification purpose only, we use the output of the second fully-connected layer as our face feature, which is a 4096-dim vector for each input face.

Before extracting descriptors, the face region proposed by the face detector is further geometrically normalized to reduce the scale uncertainty in the detector output and the effect of pose variation, e.g. in-plane rotation. An affine transformation is estimated which transforms the located facial feature points to a canonical set of feature positions. The affine transformation defines an ellipse which is used to geometrically normalize the circular region around each feature point from which local appearance descriptors are extracted [17].

3.3 Verification Metric Learning

A discriminatively learned metric was used on top of the CNN features. First, the 4096 face feature is still computationally demanding for real-time face verification. Second, since the VGG face model is often not particularly trained for verification purpose. Adding a metric learning could allow to re-tune a generically learned CNN face model towards the verification purpose while avoiding the expensive back propagation procedure.

For each face we extract n pairs of distances. The Mahalanobis distance is defined as $(a_i - y)^T M (a_i - y)$ where the positive semi-definite matrix M is a distance metric that we want to learn from training data. Simply M can be the inverse of the covariance matrix between a_i and y. Then, if M is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. Inspired by the definition of covariance matrix, we would like to decompose M into ww^T where w is a projection matrix. Then, our objective is to

$$\min_{w} \sum_{i} (a_i - y)^T w w^T (a_i - y) \tag{2}$$

and equivalently

$$\min_{w} \sum_{i} (w^T a_i - w^T y)^T (w^T a_i - w^T y). \tag{3}$$

where the projection matrix w characterizes a discriminative low-dimensional subspace. Thus, applying a distance metric also serves as dimension reduction. A Mahalanobis distance implicitly corresponds to computing the Euclidean distance after the linear projection of the data. Once w is learned from training data, then we fix w for testing. In the following, we will define a supervised learning objective which minimizes the distances of genuine pairs and maximizes the distances of impostor pairs. The loss function is a weighted combination of the two aspects.

Given two sequences x_i and y_m, we have two set of training samples $\left\{ \left((x_i, x_j), 1\right), \left((y_m, y_n), 1\right) \right\}_+$ and

$\Big\{\big((x_i, y_m), 0\big)\Big\}_-$. We propose Supervised Multi-View Canonical Component Analysis (SMVCCA), which can be formulated as the following optimization problem.

$$\begin{aligned} \max_{\mathbf{W}} \quad & trace(\mathbf{W}^T\mathbf{C}\mathbf{W}) \\ s.t \quad & \mathbf{W}^T\mathbf{C}_{d_{xy}}\mathbf{W} = \mathbf{I} \\ & \mathbf{W}_{(1)}^T\mathbf{C}_{d_{xy}}^{(11)}\mathbf{W}_{(1)} = \ = \mathbf{W}_{(K)}^T\mathbf{C}_{d_{xy}}^{(KK)}\mathbf{W}_{(K)} \\ & = \mathbf{W}_{(y)}^T\mathbf{C}_{d_{xy}}^{(yy)}\mathbf{W}_{(y)}, \end{aligned} \tag{4}$$

where $\mathbf{C} = \mathbf{C} + \mathbf{C}_{d_{xy}} = \mathbf{Z}\mathbf{Z}^T$, $\mathbf{Z} = [\mathbf{X}\ \mathbf{Y}]$. The covariance $\mathbf{C}_{d_{xy}}$ is only a normalization term. Therefore, we can neglect it, and correspondingly, change Eq. (4) to be:

$$\begin{aligned} \max_{\mathbf{W}} \quad & trace(\mathbf{W}^T\mathbf{C}\mathbf{W}) \\ s.t \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I} \\ & \mathbf{W}_{(1)}^T\mathbf{W}_{(1)} = \ = \mathbf{W}_{(K)}^T\mathbf{W}_{(K)} = \mathbf{W}_{(y)}^T\mathbf{W}_{(y)}. \end{aligned} \tag{5}$$

It has been verified [**?**] that Eq. (5) can be rewritten as a least-squares problem:

$$\begin{aligned} \min \quad & \|\mathbf{Z} - \mathbf{W}\mathbf{H}\|_F^2 \\ s.t \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I} \\ & \mathbf{W}_{(1)}^T\mathbf{W}_{(1)} = \ = \mathbf{W}_{(K)}^T\mathbf{W}_{(K)} = \mathbf{W}_{(y)}^T\mathbf{W}_{(y)}. \end{aligned} \tag{6}$$

where $\mathbf{H}$ is the coefficient matrix, and $\mathbf{W}$ is the basis matrix. Following this formulation, group-sparse nonnegative SMVCCA is proposed as:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{Z} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{H}\|_F^2 + \beta\sum_{k=1}^{K}\|\mathbf{W}_{(k)}\|_{1,q}, \\ s.t. \quad & \forall\|\mathbf{w}_i^{(k)}\|^2 \le 1, k = 1, \cdots, K, \\ & \forall\|\mathbf{w}_i^{(y)}\|^2 = 1, \ \ i = 1, \cdots, r, \\ & \mathbf{H} \ge \mathbf{0},\ \mathbf{W} \ge \mathbf{0}. \end{aligned} \tag{7}$$

Here $\mathbf{H} \ge \mathbf{0}, \mathbf{W} \ge \mathbf{0}$ are the nonnegative constraints on both the basis and coefficients. $\|\mathbf{w}_i^{(k)}\|^2 \le 1$ is a convex relaxation of each $\|\mathbf{w}_i^{(k)}\|^2 = 1$, which ensures that the correlations are normalized. The penalty $\|\mathbf{H}\|_F^2$ is to avoid arbitrarily large $\mathbf{H}$. Parameters α, β control the relative influence of each penalty term. The group sparsity penalty on the K-view canonical basis $\mathbf{W}_{(1:K)}$ is the $\ell_{1,q}$-norm: $\beta\sum_{k=1}^{K}\|\mathbf{W}_{(k)}\|_{1,q}$. The most general value for q is 2 or ∞. In particular, in order to promote sparsity on the feature views, we adopt $q = \infty$. Each $\ell_{1,q}$-norm here is defined by

$$\|\mathbf{W}\|_{1,q} = \sum_{i=1}^{r}\|\mathbf{w}_i\|_q = \|\mathbf{w}_1\|_q + \cdots \|\mathbf{w}_r\|_q, \tag{8}$$

which is the sum of vector ℓ_q-norms of its columns. Such $\ell_{1,q}$-norm is used to promote that canonical basis matrices $\mathbf{W}_{(1)}, \cdots \mathbf{W}_{(K)}$ contain as many zero columns as possible, which corresponds to only the correlations of the non-zero feature views are maximized in certain canonical vector space. Notice that there is sparsity penalty on the label view and we retain the normalization $\|\mathbf{w}_i^{(y)}\|^2 = 1$. This is because all feature views are expected to be highly correlated with the label view in canonical space.

In fact, the $\ell_{1,q}$ penalty is related to the constraints $\forall\|\mathbf{w}_i^{(k)}\|^2 \le 1$. As features are sparser, their ℓ_2-norm will be naturally smaller. Therefore, we change $\forall\|\mathbf{w}_i^{(k)}\|^2 \le 1$ to be $\forall\|\mathbf{w}_i^{(k)}\|^2 \le 1 - \beta$, and, $\forall\|\mathbf{w}_i^{(y)}\|^2 = 1 - \beta$. Finally, group-sparse nonnegative SMVCCA is expressed as:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{Z} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{H}\|_F^2 + \beta\sum_{k=1}^{K}\|\mathbf{W}_{(k)}\|_{1,q}, \\ s.t. \quad & \forall\|\mathbf{w}_i^{(k)}\|^2 \le 1 - \beta, k = 1, \cdots, K, \\ & \forall\|\mathbf{w}_i^{(y)}\|^2 = 1 - \beta, \ \ i = 1, \cdots, r, \\ & \mathbf{H} \ge \mathbf{0},\ \mathbf{W} \ge \mathbf{0}. \end{aligned} \tag{9}$$

Optimization

Problem (9) is biconvex with respect to $\mathbf{W}$ and $\mathbf{H}$. Since the sparsity penalties are non-smooth, we develop an alternative optimization scheme based on the block coordinate descent (BCD) method. In particular, the scheme bases on the vector-block BCD method, in which one column of a factor matrix is updated at each step fixing all other values. It has been proven that such vector-block BCD often outperforms matrix-block BCD method.

The overall vector-block BCD method is summarized in the following. The inner loop are the vector-block optimization of $\mathbf{H}$, $\mathbf{W}_{(1:K)}$ and $\mathbf{W}_{(y)}$, respectively.

The solution of the subproblem is given in a closed form:

$$\mathbf{h}_{i\cdot} \leftarrow \left[\frac{\mathbf{w}_i^T\mathbf{R}_i}{2\alpha + \|\mathbf{w}_i\|^2}\right]_+. \tag{10}$$

where $[]_+$ denote the element-wise projection operator to nonnegative numbers.

The subproblem is easily seen to be equivalent to

$$\begin{aligned} \min \quad & \|\mathbf{w}_i^{(k)} - \frac{\mathbf{R}_i^{(k)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\|_2^2 + \frac{\beta}{\|\mathbf{h}_{i\cdot}\|^2}\|\mathbf{w}_i^{(k)}\|_q, \\ s.t. \quad & \mathbf{w}_i^{(k)} \ge \mathbf{0}, \|\mathbf{w}_i^{(k)}\|^2 \le 1 - \beta. \end{aligned} \tag{11}$$

Given the nonnegative constraint, the following problem retains the minimum of problem (11):

$$\min_{\|\mathbf{w}_i^{(k)}\|^2 \le 1-\beta} \|\mathbf{w}_i^{(k)} - \left[\frac{\mathbf{R}_i^{(k)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\right]_+\|_2^2 + \frac{\beta}{\|\mathbf{h}_{i\cdot}\|^2}\|\mathbf{w}_i^{(k)}\|_q. \tag{12}$$

As $q = \infty$, based on the theory of Fenchel norm duality [**?**], the dual form of problem (12) is

$$\begin{aligned} \min \quad & \|\mathbf{w}_i^{(k)} - \left[\frac{\mathbf{R}_i^{(k)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\right]_+\|_2^2 + \|\mathbf{w}_i^{(k)}\|_q, \\ s.t. \quad & \|\mathbf{w}_i^{(k)}\|_1 \le \frac{\beta}{\|\mathbf{h}_{i\cdot}\|^2},\ \|\mathbf{w}_i^{(k)}\|^2 \le 1 - \beta, \end{aligned} \tag{13}$$

which can be solved by first using the method, and then normalization such that $\|\mathbf{w}_i^{(k)}\|^2 \le 1 - \beta$. Once the minimizer

$\mathbf{w}_i^{(k)*}$ of problem (12) is computed, the optimal solution for problem (11) is found as

$$\mathbf{w}_i^{(k)**} = \left[\frac{\mathbf{R}_i^{(k)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\right]_+ - \mathbf{w}_i^{(k)*}. \quad (14)$$

The solution of the subproblem is to first minimize the cost function

$$\mathbf{w}_i^{(y)} \leftarrow \left[\frac{\mathbf{R}_i^{(y)}\mathbf{h}_{i\cdot}^T}{\|\mathbf{h}_{i\cdot}\|^2}\right]_+, \quad (15)$$

then normalize the solution by

$$\mathbf{w}_i^{(y)} \leftarrow \frac{\mathbf{w}_i^{(y)}}{\|\mathbf{w}_i^{(y)}\|}\sqrt{1-\beta}. \quad (16)$$

3.4 Recognition at Run Time

A current strategy which works reasonably well in practice is to claim the person at registration and the person in testing are the same one, as long as there have been τ accumulated time stamps with the respective smallest distance below a threshold. Let us write $A = [a_1, ..., a_n]$ and then we have a problem of

$$y = Ax \quad s.t. \quad \|x\|_0 = 1 \quad (17)$$

at each time step, where x is an indicator vector. Surely, only accounting the reference images with the smallest distance is from experiences yet heuristic. If we hope to account no more than a sparsity s reference images, then we have

$$y = Ax \quad s.t. \quad \|x\|_0 \leq s \quad (18)$$

Once $\mathbf{W}$ is determined, given a test sample $\mathbf{x}$ and its label $\mathbf{y}$, the inference to its corresponding coefficient vector $\mathbf{h}$ is

$$\min_{\mathbf{h}\geq 0} \frac{1}{2}\|\mathbf{z} - \mathbf{W}\mathbf{h}\|_F^2 + \alpha\|h\|_F^2 \quad (19)$$

where $\mathbf{z} = [x^T y^T]^T$. It is easy to show that this problem has the following closed-form solution:

$$\mathbf{h} = \left[(\mathbf{W}^T\mathbf{W} + 2\alpha\mathbf{I})^{-1}\mathbf{W}^T\mathbf{z}\right]_+ \quad (20)$$

4 Experiments

We test our network on the UNBC-McMaster Shoulder-Pain dataset. It contains 200 videos of 25 patients who repeatedly raise their arm (feeling pain) and then put it down (pain released). All frames per video are labeled with a ground-truth pain score. Now, our task is to fit our predicted score into the ground truth during testing. Due to our network's nature of our network that there is no temporal modeling at all, we test each frame separately as an face image. As a result, it does not affect the MSE even if we mix all videos of one patients and shuffle all images. However, the x-axis Fig. 4 indexes the frames in the time order and videos in the directory alphabetical order.

We run leave-one-out cross validation 25 times. Each time, the videos of one patient are reserved for testing. All the other videos are used to train the deep regression network. As a result, we train the network 25 times. Even with pre-trained models being available from the previous round, We simply re-train the network using the current eligible training data. There is no sharing of network weights across different rounds of cross validation. In the end, the performance is summarized in Table 1.

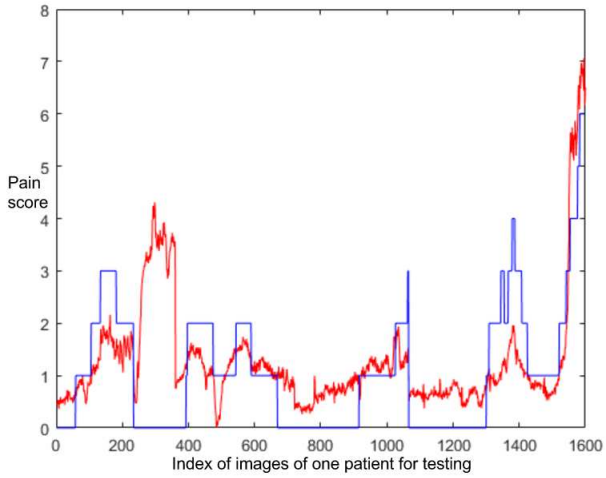

Figure 4: Our regression network with the center loss.

Methods	MAE	MSE	PCC
smoothL1 + ReLU	0.456	0.936	0.541
smoothL1 + sigmoid	0.416	1.060	0.524
smoothL1 + softmax + sigmoid	0.394	1.039	0.485
L1 + centerloss + sigmoid	***0.389***	0.820	0.603
smoothL1 + L1 centerloss + sigmoid	0.456	***0.804***	***0.651***
smoothL1 + L2 centerloss + sigmoid	0.435	0.816	0.625
OSVR-L1 (CVPR16) [22]	1.025	N/A	0.600
OSVR-L2 (CVPR16) [22]	0.810	N/A	0.601
RCNN (CVPR16w) [23]	N/A	1.54	***0.65***

Table 1: Performance our various versions of our regression network on the Shoulder-Pain dataset for automated assessment of the pain level (*i.e.*, pain expression intensity). MAE is short for mean abosolute error deviated from the ground-truth label over all frames per video. MSE is mean squared error which measures the curve fitting degree. PCC is Peason correlation coefficient which measures the curve trend similarity (the larger, the better).

5 Conclusion

In summary, the main contribution of this paper is a video-versus-video consensus with discriminative metric learning on the fly, which is verified in a working system for the crowd monitoring system in the hospital.

References

[1] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[2] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[3] Chopra, Sumit, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification." Computer Vision and Pattern Recognition,

2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.
[4] Huang, Gary B., et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Vol. 1. No. 2. Technical Report 07-49, University of Massachusetts, Amherst, 2007.
[5] H. Rowley, S. Baluja and T. Kanade. Neural network-based face detection. In IEEE CVPR, 1996.
[6] M. Szarvas, A. Yoshizawa, M. Yamamoto and J. Ogata. Pedestrian detection with convolutional neural networks. In Intelligent Vehicles Symposium, 2005.
[7] D. Pomerleau. ALVINN: An Autonomous Land Vehicle in a Neural Network. In NIPS, 1989.
[8] Y. LeCun et al. Backpropagation Applied to Handwritten Zip Code Recognition.Neural Compuation, 1(4), 1989.
[9] A. Krizhevsky, I. Sutskever, GE Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.
[10] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." Proceedings of the British Machine Vision 1.3 (2015): 6.
[11] Liu, Chengjun, and Harry Wechsler. "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition." Image processing, IEEE Transactions on 11.4 (2002): 467-476.
[12] Ahonen, Timo, Abdenour Hadid, and Matti Pietikainen. "Face description with local binary patterns: Application to face recognition." Pattern Analysis and Machine Intelligence, IEEE Transactions on 28.12 (2006): 2037-2041.
[13] Q. Cao, Y. Ying and P. Li. Similarity Metric Learning for Face Recognition. [14] M. Guillaumin, J. Verbeek and C. Schmid. Is that you? Metric Learning Approaches for Face Identification.
[15] J. Hu, J. Lu and Y-P Tan. Discriminative Deep Metric Learning for Face Verification in the Wild. In IEEE CVPR, 2014.
[16] Sun, Yi, et al. "Deep learning face representation by joint identification-verification." Advances in Neural Information Processing Systems. 2014.
[17] M. Everingham, J. Sivic, and A. Zisserman. hello! my name is... buffy - automatic naming of characters in tv video. In BMVC, 2006.
[18] Ojala, Timo, Matti Pietikinen, and Topi Menp. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Trans. PAMI 24.7 (2002): 971-987.
[19] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, Xilin Chen: A Benchmark and Comparative Study of Video-Based Face Recognition on COX Face Database. IEEE Trans. Image Processing 24(12): 5967-5981 (2015)
[20] Zagoruyko, Sergey, and Nikos Komodakis. "Learning to compare image patches via convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
[21] Wolf, Lior, Tal Hassner, and Itay Maoz. "Face recognition in unconstrained videos with matched background similarity." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.
[22] Rui Zhao, Quan Gan, Shangfei Wang,and Qiang Ji. Facial expression intensity estimation us-ing ordinal information. InThe IEEE Conference on Com-puter Vision and Pattern Recognition (CVPR), June 2016.
[23] Jing Zhou, Xiaopeng Hong, Fei Su, andGuoying Zhao. Recurrent convolutional neural networkregression for continuous pain intensity estimation invideo.arXiv preprint arXiv:1605.00894, 2016.