LEARNING EXPRESSION KERNELS FOR FACIAL EXPRESSION INTENSITY ESTIMATION

Chia-Te Liao, Hui-Ju Chuang, and Shang-Hong Lai

Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

Although many studies of facial expression analysis have been conducted, most previous works indeed focused on expression recognition. Different from previous works, this paper proposes a novel approach to learn the expression kernel for facial expression intensity estimation. The solution involves first aligning the optical flow to a neutral face to reduce inter-person variations in facial geometry, followed by solving an optimization problem with the ordinal ranking of expression intensities in temporal domain as constraints. Extensive experiments on the Cohn-Kanade database manifest that using the learned expression kernels leads to superior performance than the previous methods for facial expression intensity estimation.

Index Terms—Facial expression analysis, expression intensity estimation, quadratic programming.

1. INTRODUCTION

Facial expression is a dynamic process from onset to apex, which is about expression intensity variation in temporal domain. To analyze the expression dynamics, only classifying expressions into basic categories is insufficient for obtaining the in-depth understanding of human emotion. Also because estimating the expression intensity is not a hard decision problem, the conventional classification methods are unsuitable as illustrated in Figure 1. On the other hand, the regression methods can not be used either because we can not have ground-truth of absolute intensities. These issues make intensity estimation a challenging task. So far, the estimation of facial expression intensity has only been studied by a few previous works (e.g. [1, 2]).

In the literature, several previous works attempt to recognize fine-grained changes in facial expression based on the Facial Action Coding System (FACS) [3], which decomposes a facial expression into a number of specific action units (AUs). Clearly, for face images of the same expression, the local motion in some specific facial regions is similar because the same AUs are shared. However, it is still an open problem to automatically label the intensity of AUs due to FACS does not give a clear definition for AU's intensity level. In view of this, we aim to build a function which simulates the combination of AUs and outputs an estimation of the expression intensity. This paper proposes a learning approach to building kernels to be used with kernel machines for facial expression intensity estimation. We formulate the learning task as a multi kernel learning (MKL) problem, where the aligned intra-person optical flow is used to alleviate the inter-person variations. The learned kernel can better measure the similarity between face images as human expectations for an expressional face, thus providing advantages for facial expression analysis. It is ready to be used in conjunction with kernel machines, such as SVM, for expression classification. The rest of this paper is organized as follows: In Section 2, we review some related works about kernel learning. Section 3 describes the details of the learning framework, which consists of the expression flow computation and the optimization problem formulation for learning the facial expression kernels. Section 4 gives experimental results for the proposed application. Finally we concluded this paper in the last section.

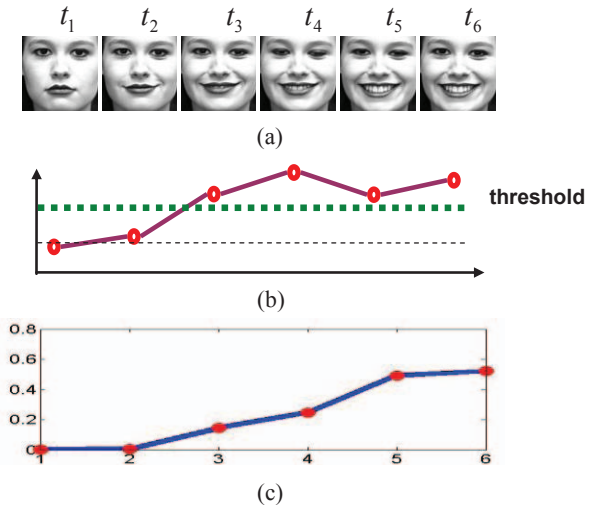

Figure 1. (a) A happiness image sequence with increasing expression intensities, **(b)** Output of an expression recognition classifier, and **(c)** Intensity estimation results by using the proposed expression kernel.

2. MULTIPLE KERNEL LEARNING

For kernel machines, the underlying kernel $k(\bullet,\bullet,|\theta)$, where θ denotes the type of the kernel and its parameter values, plays a very crucial role for the performance of the kernel machines. In crafting appropriate kernels, researches attempt to find an optimal way to linearly combine M given kernels to obtain a stronger kernel function $\hat{k}$, i.e.

$$\hat{k}(\bullet,\bullet)=\sum_{j=1}^{M}\beta_j k(\bullet,\bullet\,|\,\theta_j),\beta_j\geq 0. \quad (1)$$

Connecting the ensemble kernel with a kernel machine for binary class data $\{(x_i,y_i)\in\pm 1\}_{i=1}^{N}$, for example, will result in the following formulation:

$$f(x)=\sum_{i=1}^{N}\alpha_i\hat{k}(x,x_i)+b=\sum_{i=1}^{N}\alpha_i\sum_{j=1}^{M}\beta_j k(\bullet,\bullet\,|\,\theta_j)+b \quad (2)$$

where b is a constant bias and α_i are coefficients determined from training. Intuitively the base kernels with higher β_j values are deemed more useful, thus determining the optimal coefficients $\{\beta_1,...,\beta_M\}$ corresponds to finding appropriate weights for best combining the M feature representations in terms of M kernel matrices. Researchers have achieved many successful results with different kernel machines by using the MKL approach (e.g. [4-6]).

3. PROPOSED EXPRESSION KERNEL

This section details the proposed approach to building expression kernels for expression intensity estimation.

3.1. Optical flow computation and normalization

Because optical flow preserves good information of the correlation between two images, we compute it as the features of face images used in our expression kernel for expression analysis. In this work, we employed the constrained optical flow algorithm proposed in [7] to compute the optical flow between two face images. However, different people have distinct facial geometry; a normalization procedure is required to eliminate the inter-person variation. Let us denote $EX_p^{(Ei)}$ as an expression image of expression *Ei*, NE_p is the neutral face image of person *p*, and *OF*( ) is a directional operator defined for the optical flow computation with two images. The normalization procedure is started from a global neutral face NE_0 to obtain the inter-person optical flow $OF_{\text{inter},p}=OF(NE_0,NE_p)$ and the overall optical flow $OF_{all}^{(Ei)}=OF(NE_0,EX_p^{(Ei)})$. The intra-person optical flow can be computed for each facial expression image by subtracting the overall optical flow from the inter-person optical flow as summarized in Figure 2.The input optical flows are thus represented with the same geometry of NE_0.

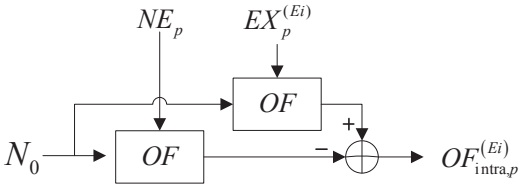

Figure 2. Align expression flows to NE_0

Given an expressional face image, we compute its intra-person expression flow as the representation and learn the kernels for expression analysis in the next subsection.

3.2. Learning an expression kernel

To estimate the expression intensity, a measure on locations where the associated AUs occur is required. Rather than using the hand-crafted rules, here we formulate a constrained optimization problem based on MKL to learn an expression kernel for intensity estimation. Let $\hat{k}_{\text{exp}}$ be a kernel function defined to measure the similarity of two expression optical flows given by:

$$\hat{k}_{\text{exp}}(OF_{\mathbf{a}},OF_{\mathbf{b}})=\sum_j w_j k_j\left(OF_{\mathbf{a}},OF_{\mathbf{b}}\right), \quad (3)$$

where k_j is a dot-product kernel defined on two vectors corresponding to the j^{th} component of expression flows $OF_{\mathbf{a}}$ and $OF_{\mathbf{b}}$, denoted by $(OF_{\mathbf{a}})_j$ and $(OF_{\mathbf{b}})_j$:

$$k_j\left(OF_{\mathbf{a}},OF_{\mathbf{b}}\right)=\left\langle\frac{(OF_{\mathbf{a}})_j}{\left\|(OF_{\mathbf{a}})_j\right\|},\frac{(OF_{\mathbf{b}})_j}{\left\|(OF_{\mathbf{b}})_j\right\|}\right\rangle \quad (4)$$

Our aim is to derive an expression kernel $\hat{k}_{\text{exp}}$ by MKL whose kernel values between expression flows of similar expression intensities are larger than those of different expressions or intensities. Given a video sequence, suppose the intensity of each frame increases from the start state to the apex with time *t*. As the ranking-based framework [2], we use the pair-wise ordinal relationship along the temporal domain to provide the ordinal constraints in building $\hat{k}_{\text{exp}}$. Let $OF_{apex}^{(Ei)}$ be the expression template of expression *Ei* computed by averaging the optical flows of apex images of all available sequences. Since the expression kernel $\hat{k}_{\text{exp}}$ measures similarity in the expression intensity space, the kernel value between $OF_{apex}^{(Ei)}$ and an expression flow with stronger intensity, e.g. $OF_t^{(Ei)}$, should be larger than that between $OF_{apex}^{(Ei)}$ and $OF_{t-1}^{(Ei)}$, i.e..

$$\hat{k}_{\text{exp}}(OF_t^{(Ei)},OF_{apex}^{(Ei)})>\hat{k}_{\text{exp}}(OF_{t-1}^{(Ei)},OF_{apex}^{(Ei)})>\cdots$$
$$>0>\hat{k}_{\text{exp}}(OF_t^{(Ej)},OF_{apex}^{(Ei)}) \quad (5)$$

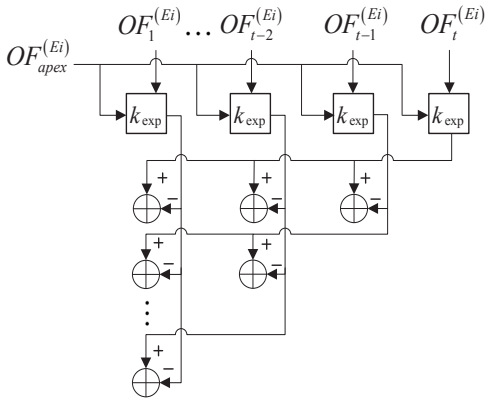

Figure 3. Illustration of constraint formulation according to (5). Each $\oplus$ represents an ordinal constraint in (6).

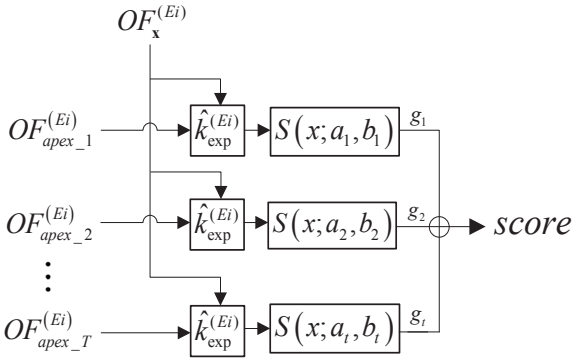

Figure 4. Estimating the facial expression intensity for an input face image *x* with the learned expression kernel.

According to the ordinal ranking of expression intensities, a constrained optimization problem is formulated. The solution clearly leads to the optimal weighting configuration $\mathbf{w}=[w_1, w_2, .., w_n]$ from the training data. It is formulated by maximizing the weighted gap between the kernel values, while minimize the training errors. The constrained optimization problem for the optimal **w** is given as follows:

$$\begin{aligned} &\min_{\mathbf{w},\xi,\rho} \alpha\left(\frac{1}{2}\mathbf{w}^T Q\mathbf{w}\right) - v\rho + C\sum_{i=1}^{m}\xi_i \\ &s.t \quad \sum_{j=1}^{n} w_j k_j\left(OF_{\mathbf{a}}^{(Ei)}, OF_{apex}^{(Ei)}\right) - \sum_{j=1}^{n} w_j k_j\left(OF_{\mathbf{b}}^{(Ei)}, OF_{apex}^{(Ei)}\right) \geq \rho - \xi_i, \\ &\quad \|\mathbf{w}\| = 1, \ \xi_i \geq 0, \quad \rho \geq 0, \ \ i = 1, ..., m \end{aligned} \tag{6}$$

where Q is a matrix corresponding to the 2D Laplacian operator that imposes spatial smoothness of **w**, and m is the number of all possible combinations of $\left(OF_{\mathbf{a}}^{(Ei)}, OF_{\mathbf{b}}^{(Ei)}\right)$ derived from training data. $\left(OF_{\mathbf{a}}^{(Ei)}, OF_{\mathbf{b}}^{(Ei)}\right)$ denotes a pair of expression flows where $OF_{\mathbf{a}}^{(Ei)}$ is labeled with stronger intensity than $OF_{\mathbf{b}}^{(Ei)}$ according to their temporal order within a facial expression image sequence. The slack variables ξ_i for the i^{th} constraint is introduced to tolerate that example violates the constraint. In order not to obtain the trivial solution where all ξ_i take on large values, we penalize them in the objective function with a penalty parameter C. The parameter α is to balance the smoothness term (on the left-hand side) and data term (on the right-hand side), and v is a parameter to control the significance of separation ρ for training examples. Also, the coefficients in **w** should be able to enlarge ρ as wide as possible to obtain better generalization capability.

The resulted optimization problem is a convex quadratic programming (QP) problem, which has a global optimum and can be easily solved by using a standard QP solver. Let S denote the union set of all subsets of expression optical flows of different expression types, i.e. $S=\{S_1, S_2, \ldots, S_{|S|}\}$. By using the one-against-rest approach to assume the k^{th} expression as the target expression, i.e. $S^+ = S^k$, and using the others as negative examples, i.e. $S^- = S - S^k$, we train an expression kernel for each expression k by solving the optimization problem in Equation (6). Figure 3 illustrates the constraint formulation procedure for more clarity.

4. APPLICATION TO EXPRESSION INTENSITY ESTIMATION

Since a kernel function can be viewed as a similarity function encoding the prior knowledge, it is possible to give a face image **x** a relative intensity measure through comparing it with a typical expression template for reference. The problem of estimating an expression intensity measure is thus reduced to computing the similarity of two expression flows using kernel $\hat{k}_{\exp}^{(Ei)}$. As shown in Figure 4, we give the expression intensity score to an input face image x through computing $\hat{k}_{\exp}\left(OF_x^{(Ei)}, OF_{apex}^{(Ei)}\right)$ with the scoring function:

$$score_{Ei}(x) = \sum_{t=1}^{T} g_t \frac{1}{1+\exp(-a_t \hat{k}_t^{(Ei)}(OF_x^{(Ei)}, OF_{apex\_t}^{(Ei)}) + b_t)}. \tag{7}$$

Because there are several styles to express the same emotion for different people (for example, the happiness can be shown by either smiling or laughing loudly), the expression intensity score is the weight-averaged score using T templates of expression Ei. A sigmoid function with parameter a_t and b_t is here used to scale the kernel function output to [0,1], and g_t is the prior of style t. $OF_x^{(Ei)}$ is un-normalized here to receive stronger response for larger facial motion, and $\hat{k}_t^{(Ei)}$ is trained using the clustered examples with the t^{th} template. Parameters g_t, a_t, and b_t can be estimated via least-squares fitting the score function in eq. (7) to the intensity-labeled data.

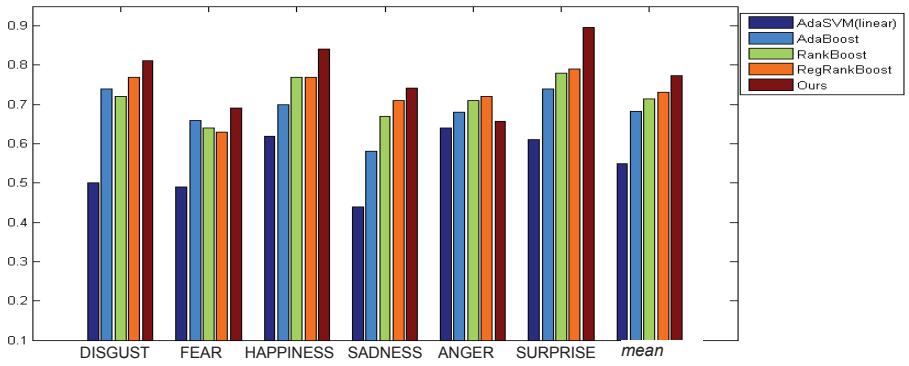

Figure 5. Comparison of relative accuracy measures for different methods [2] on the testing set

In Figure 5, we give experimental results by using the CMU Cohn-Kanade database [8]. The *Relevant Accuracy* (RA) [9] is used for evaluating the performance of intensity estimation with the definition:

$$RA = \frac{\#\ of\ correctly\ ranked\ relevant\ pairs}{\#\ of\ all\ relevant\ pairs} \qquad (8)$$

Given a sequences with n frames, there are C_2^n relevant pairs to test the accuracy with the ground truth of ordinal relationship along time. An interval of 3 frames is applied to rebuilt image sequences as [2] since the variations of consecutive images are too subtle to be distinguished. We also compare the RA measure in Figure 5 by using the proposed intensity estimation algorithm with results by using *AdaSVM* [10], *AdaBoost* [11], *RankBoost* and *RegRankBoost* [2] algorithms with the haar-like features. Although the RankBoost and RegRankBoost applied the similar ordinal ranking concept in the learning procedure, our intensity estimation results demonstrates the higher accuracy for all expressions except the anger expression. This is because using the AdaBoost-based approaches with the Haar-like features is prone to select weak learners of outlier local features, such as wrinkles. In contrast, our kernel is completely determined by the associated expression optical flows. However, for anger expression the proposed kernel-based method does not perform better than others partly because the motion of compressing lips in anger expression is not easily detected from the expression flows.

5. CONCLUSIONS

This paper proposed a novel approach to learn the expression kernel for facial expression intensity estimation. The expression motion of a facial expression image is represented by the optical flow, while a normalization strategy is performed to reduce the inter-person variations in facial geometry. An MKL-based learning approach is then used to learn the expression kernels, where the weighting coefficients are determined by solving a constrained optimization problem. Extensive experiments on the Cohn-Kanade facial expression database showed the advantages of the proposed expression kernel on expression intensity estimation. In the future, we plan to extend this expression kernel by including the temporal and appearance information to further improve the accuracy for facial video analysis.

6. REFERENCES

[1] J. J.-J. Lien, T. Kanade, J.F. Cohn, C.C. Li, and A.J. Zlochower, "Subtly different facial expression recognition and expression intensity estimation," *Proc. CVPR*, 1998, pp. 853-859.

[2] P. Yang, Q. Liu, and D. N. Metaxas, "RankBoost with L_1 regularization for facial expression recognition and intensity estimation," *Proc. ICCV*, 2009, pp. 1018-1025.

[3] P. Ekman and W.V. Friesen, *Facial action coding system (FACS): Manual*, Consulting Psychologists Press, 1978.

[4] S. Sonnenburg, G. R¨atsch, C. Sch¨afer, and B. Sch¨olkopf, "Large scale multiple kernel learning," *JMLR*, 7, 2006, pp. 1531-1565.

[5] M. Gonen and E. Alpaydin, "Localized multiple kernel learning," *Proc. ICML*, 2008, pp. 352-359.

[6] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," *Proc. ICML*, 2007, pp. 775-782.

[7] C.K. Hsieh, S.H. Lai, and Y.C. Chen, "Expression-invariant face recognition with constrained optical flow warping," IEEE Trans. Multimedia, vol. 11, no. 4, pp. 600-610, 2009.

[8] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proc. FG'00*, pp. 46-53.

[9] H. A. Bangpeng Yao and S. Lao, "Logit-rankboost with pruning for face recognition," *Proc. FG'08*, pp.1-8.

[10] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *J. Image and Vision Computing*, 2006, pp.615-625.

[11] S. Koelstra and M. Pantic, "Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics," *Proc. FG'08*, pp.1-8.