ONE SHOT EMOTION SCORES FOR FACIAL EMOTION RECOGNITION

Albert C. Cruz, B. Bhanu, and N. S. Thakoor

Center for Research in Intelligent Systems, University of California, Riverside
Winston Chung Hall 216, Riverside, CA, 92521-0425, USA

ABSTRACT

Facial emotion recognition in unconstrained settings is a difficult task. They key problems are that people express their emotions in ways that are different from other people, and, for large datasets, there are not enough examples of a specific person to model his/her emotion. A model for predicting emotions will not generalize well to predicting the emotions of a person who has not been encountered during the training. We propose a system that addresses these issues by matching a face video to references of emotion. It does not require examples from the person in the video being queried. We compute the matching scores without requiring fine registration. The method is called *one-shot emotion score*. We improve classification rate of interdataset experiments over a baseline system by 23% when training on MMI and testing on CK+.

Index Terms— Emotion recognition, similarity measures

1. INTRODUCTION

Emotion recognition has applications in: human-computer interaction, video games, medicine, and deception detection. In facial emotion recognition, a video of a human face is captured. Computer algorithms must detect facial expressions and infer their underlying emotional state. There are two key problems to state-of-the-art methods:

1) Uniqueness of expressions. A person's expressions and gestures are so unique to a person that they can be used for identification [6]. An example of different expressions of happiness is given in Figure 1.

2) Insufficient examples. Unconstrained facial emotion recognition is *person-independent*. The system must predict emotions of new individuals that are not in the training data. Because of uniqueness of expressions, in testing, there are insufficient examples to properly describe the emotion of a new individual. The top approach for the Facial Emotion Recognition and Analysis grand challenge [5,7] achieved 96% with person-dependent experiments, but the performance dropped 21% when conducting person-independent experiments. The performance of an approach for the Audio/Visual Emotion Challenge [8] dropped 19% between training and testing.

We address each of these technical challenges with *learning with side information*. Conventionally, a method projects all the samples into a decision space based on their feature vector. However, we predict the emotion of a video by matching the video to background information and a target emotion reference. A score is computed that determines if the video is more similar to the target emotion, or other emotions (*background information*). We are inspired by the approach in Wolf et al. [9] applied to Labeled Faces in the Wild (LFW) [10], where some individuals have only a single sample in the data. A summary of related work is given in Table 1.

We contribute the following to the state-of-the-art: We propose a method to address person-independent facial emotion recognition called one-shot emotion score. We demonstrate a method for quantifying face similarity that does not require fine registration on posed, frontal face videos. We improve the classification rate on interdataset experiments [1–5].

2. TECHNICAL APPROACH

Our system for emotion recognition consists of the following steps: (1) In training, references of emotion are created with a avatar reference image [5]. (2) When processing a video, face ROI is extracted with a boosted cascade of Haar-like features [11]. (3) The one-shot emotion scores (OSES) are computed at the midpoint emotion-apex frame in a video. They are quantified with expression energy. (4) The OSES are computed for each emotion and an initial classification is performed with the OSES as a feature for a linear SVM. (5) The final labels are estimated with decision level fusion using a linear SVM.

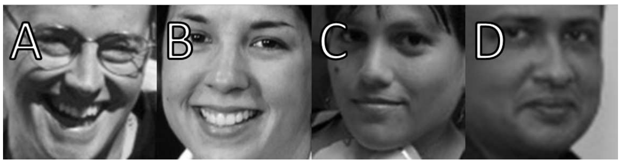

Fig. 1: Examples of different expressions of happiness. (A) A strong Duchenne smile. (B) A Pan-Am smile, sometimes considered to be disingenuous; note that, unlike the Duchenne smile, the upper cheek muscles are not used. (C) A weak Pan-Am smile. (D) Happiness without a smile.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

Table 1: Review of video-based related work. SVM: support vector machine.

Approach	Registration and Features	Classifier	Dataset
Ghanem [1]	Optical flow to track facial feature points, time series seasonality adjustment	Hidden Markov model	CK+, MMI
Li et al. [2]	Eye-point alignment, Gabor filter, active shape models	AdaBoost	CK+
Miao et al. [3]	Eye-point alignment, local binary patterns	Supervised kernel mean matching	CK+, MMI
Poursaberi et al. [4]	Gauss-Laguerre wavelets and 18 fiducial facial point features	K-NN	MMI
Yang and Bhanu [5]	Avatar image registration, local binary patterns, local phase quantization	Linear SVM	FERA, CK+
Proposed	One-Shot Emotion Score	SVM	CK+, MMI

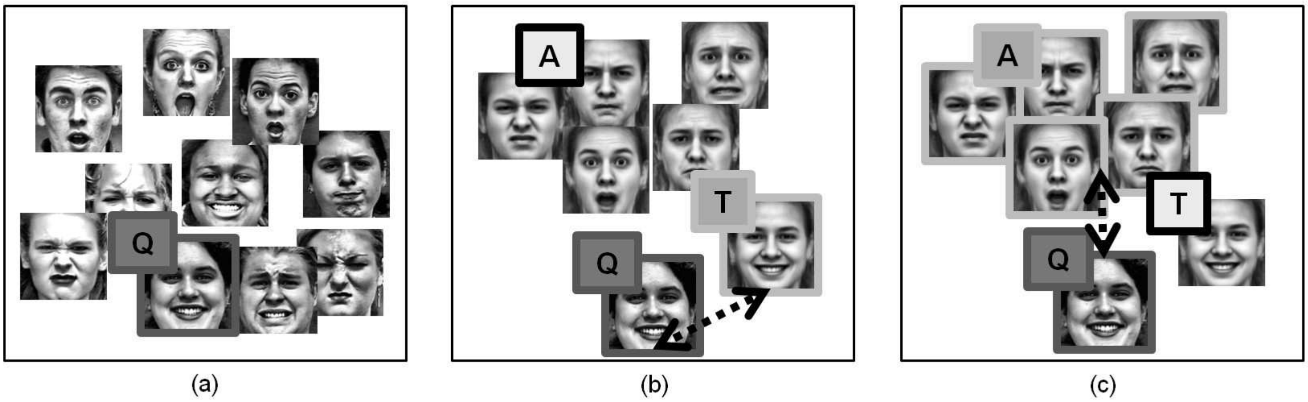

Fig. 2: (a) Approximation of a decision space by the feature vectors of $\mathcal{D}$. Because of uniqueness of expressions, there is no guarantee that similar faces are the same emotion as Q. (b-c) Approximation of a decision space created by the OSES of T_{a_j} and A. (b) We measure the distance to the target T_{a_j}, and (c) the distance to the nearest background face A. This will indicate whether Q is the emotion in T_{a_j} or A.

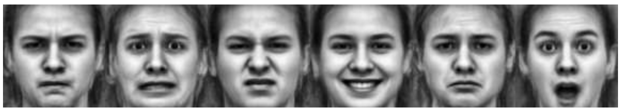

Fig. 3: Example references of characteristic emotion. (From left to right) Anger, fear, disgust, happiness, sadness and surprise.

2.1. One-Shot Emotion Scores

A query face ROI Q is submitted to the system and it has to be to determined if Q is expressing a certain emotion. Conventionally, a method would take all face images $\mathcal{D}$ including other examples of Q and project them into a decision space based on their feature vectors. Accurate classification requires enough training examples that are related to Q. This may not be possible given the technical challenges. Instead, we propose one-shot emotion scores that can be used to determine the similarity of a given face ROI to the reference of an emotion. Let Φ be the set of possible emotions: $\Phi = \{a_1, ..., a_n\}$. Let T_{a_j} be the reference of a characteristically expressed emotion, best describing a_j. Let the set of *background emotions* A be: $\Phi - \{a_j\}$. We compute two scores: one to determine how similar Q is to T_{a_j}, and one to determine how similar Q is to A. The OSES are:

$$S_{Q\rightarrow A} = \text{argmin}_{a_i \in A} M\left(Q, T_{a_i}\right) \quad (1)$$

$$S_{Q\rightarrow T_{a_j}} = M\left(Q, T_{a_j}\right) \quad (2)$$

where $M\left(.\right)$ is a function that measures the similarity between two images. These similarities can be more discriminative than a manifold trained on $\mathcal{D}$, in the presence of a limited number of training examples of Q. If Q is similar to T_{a_j}, $S_{Q\rightarrow T_{a_j}}$ should have a lower score than $S_{Q\rightarrow A}$. If a face is not similar to T_{a_j}, $S_{Q\rightarrow A}$ should have a lower score than $S_{Q\rightarrow T_{a_j}}$. A visual example is given in Figure 2.

The OSES are computed for each emotion. Equations 1 and 2 form the feature vector and an initial classification is done by SVM. There are n SVMs [12], one for each emotion, and they output decision values for each emotion in Φ according to the OSES. A second SVM fuses these values at the decision level to give a final prediction.

A problem is posed where we must generate the characteristic representation of a positively expressed emotion. The

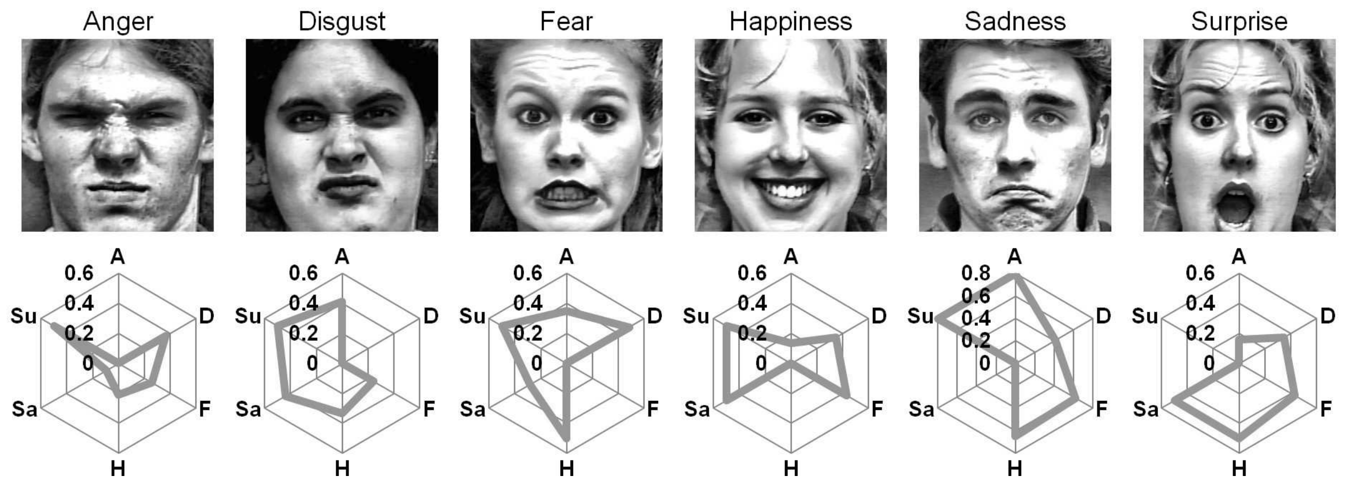

Fig. 4: Radar graphs of showing $S_{Q\to T_{a_j}}$ for six face images. Each dimension on the radar graph indicates $S_{Q\to T_{a_j}}$ of the given face image computed from the references in Figure 3. A low value indicates that the face is similar to the reference for that emotion. The values are normalized $[0, 1]$. Anger is detected as similar to sadness because of the frown, and chin movement. Disgust and happiness share deep nasolabial furrowing. During happiness, the cheeks are raised causing the appearance of squinted eyes, which is why this example of happiness is similar to anger. For all images other than surprise, surprise has the maximum value, and was clipped at .5 (.8 for sadness) to better visualize the results. This is because it has severe distortion of the face. A: anger, D: disgust, F: fear, H: happiness, Sa: sadness and Su: surprise.

avatar reference image concept [5] allows the reference of an image to be generated from multiple observations of the same image. It was originally designed to create the reference of a face for registration of face images. We generate T_{a_j} from the set of training images positively expressing a_j. Let this set be $\mathcal{D}_j$, which is a subset of $\mathcal{D}$.

Let I be an image expressing a_j. The pixel $I(x)$ is an independent observation of $T_{a_j}(x)$. We model $T_{a_j}(x)$ as a constant value that is observed in additive, normally distributed noise: $I(x) \sim N\left(T_{a_j}(x), \sigma\right)$. To estimate T_{a_j} we take the mean: $T_{a_j}(x) = \sum_{\mathcal{D}_j} I(x)$. Then, SIFT-Flow [13] aligns each face in $\mathcal{D}_j$ to T_{a_j}. This is called the avatar reference image.

The process iterates by rewarping $\mathcal{D}_j$ to the avatar reference image. A second avatar reference image is created by taking the mean of these rewarped faces. This process repeats: a new avatar reference image is estimated from the previous iteration, and all the faces would be warped to the new avatar reference image. The number of iterations is selected empirically. Examples are given in Figure 3. These reference faces were generated with two iterations. Computing the references is done in the offline training portion of the algorithm.

We quantify the scores with expression energy. It is based on SIFT-Flow [13], which has been applied to scene and face matching. We improve the cost function:

$$M_{\text{SF}}(\mathbf{w}(x)) = \sum_x \left\| s_{T_j}(x) - s_Q(x + \mathbf{w}(x)) \right\| + \alpha \sum_x \left((u(x) - \mathrm{E}(u(x)))^2 + (v(x) - \mathrm{E}(v(x)))^2 \right) \tag{3}$$

where M_{SF} is the cost function of warping Q to T_{a_j}; $\mathbf{w}(x)$ is the motion vector at pixel x matching Q to T_{a_j}; u and v are the vertical and horizontal components of $\mathbf{w}$; s_{T_j} and s_Q are the dense, per-pixel SIFT features of T_{a_j} and Q respectively; α is a weight parameter; and $\mathrm{E}(.)$ is the mean. We take M_{SF} to be M. Quantified values of $S_{Q\to T_{a_j}}$ for example faces are given in Figure 4.

Assuming that there is a SIFT feature that is matched between Q and T, an optimal warp would cause $s_T(x) - s_Q(x + \mathbf{w}(x)) = 0$. The first term sums the L_1 norm of this difference, for each pixel. The second term constrains the warp to have a small magnitude. It sums the magnitude of all the motion vectors in $\mathbf{w}(x)$.

Note that the mean value is subtracted from the second term. A whole plane translation, such as the common, slight translation errors from Viola and Jones [11], would affect each $\mathbf{w}(x)$. It would increase $M(\mathbf{w}(x))$ proportionally to the size of the image. However, because we are subtracting the mean, the effects of whole plane translation errors are removed. We can apply this algorithm directly to unregistered face ROIs on datasets where there is not much change in facial pose.

2.2. Comparison to One-Shot Similarities

Our method is inspired by one-shot similarity (OSS), but is significantly different. In OSS, Q and T are a pair of faces and the system predicts if Q and T are the same person. The background information A are junk faces that are not the same

Table 2: Summary of results and comparison to related work. D/I indicates person independent folds or person dependent folds. Bold indicates best performer. Underline indicates second best performer.

Method	# Videos	Validation	Classes	CK+	MMI	C2M	M2C
Ghanem [1]	100	D - 3 fold	Joy, anger, sadness, disgust	-	83.1	53.1	85.1
Li et al. [2]	309	I - Leave-one-subject-out	6 basic	87.4	-	-	-
Miao et al. [3]	N/A	I - Leave-one-subject-out	6 basic	-	-	55.7	-
Poursaberi et al. [4]	96	D - Leave-one-out	6 basic	-	**87.7**	-	-
Yang and Bhanu [5]	316	I - Leave-one-subject-out	6 basic	82.6	-	-	-
Proposed method	MMI: 118 CK+:296	I - Leave-one-subject-out	6 basic	**90.9**	86.4	**57.6**	**85.8**
LBP/SVM Baseline	Same as above	Same as above	Same as above	85.2	58.0	42.6	62.8

person as Q or T. A score is computed between Q and A, and T and A. We do not compute the similarity to A for both scores. In OSS, the score is quantified as an L_p norm difference between the two feature vectors. We exploit the energy of a SIFT-flow cost function to quantify the difference. OSS is for face verification and OSES are for facial emotion recognition, so the approach cannot be directly compared.

3. EXPERIMENTS

We test the person-independent generalization capability of the algorithm with interdataset experiments. There is no official testing methodology for interdataset experiments, though some works use CERT [14], which is a combination of many datasets including CK+ [15] and MMI [16]. However, CERT uses some non-public data, so we cannot use this methodology. CK+ and MMI datasets have been thoroughly addressed with intradataset experiments, with high classification rates [17, 18]. However, interdataset experiments with CK and MMI can have a classification rate as low as 42% for a baseline system. For these reasons, we conduct intradataset experiments on CK+ and MMI. The reader is referred to [15, 16] for an in depth explanation of the datasets. We classify anger, sadness, joy, disgust, surprise, and happiness. We conduct a 2-run testing validation: one where CK+ is used for training and MMI is used for testing (denoted with C2M); and MMI training with CK testing (denoted M2C). For MMI, we used all videos from all sessions of frontal face video that had an emotion label. For CK+, we used all videos that had emotion labels. We compare our results to related work and a baseline approach. The baseline approach is similar to the baseline approach in recent grand challenges [7, 19]. Face ROI is extracted, local binary patterns are extracted, and a linear SVM classifies the emotion.

For the parameters specific to SIFT-Flow: $\alpha = 510$. The avatar reference images are created for each fold, not using the face images in the testing fold. For the parameters specific to LBP [20]: the radius is 1, there are 8 neighbors, divide the image into 8×8 subregions where an LBP histogram is computed in each region, and we use uniform LBP patterns. For the parameters specific to the SVM: we use a linear SVM in all cases, and $c = 2.37$, which was empirically selected. After ROI extraction, images are resized to 200×200 and histogram equalized.

Results for inter- and intradataset experiments on CK+ and MMI are given in Table 2. The validation for intradataset experiments is leave-one-subject-out. Folds are created for each individual, to ensure person-independent generalization. Even without registration, and person independent folds, the baseline method does well on CK+. We postulate that the emotions in CK+ are more characteristic; they are more similar to the reference images. MMI is more difficult because some people in the videos are wearing glasses, and the expressions are more varied. The proposed method is the best performer for all intra- and interdataset experiments except for MMI intradataset. Poursaberi et al. [4] is the best performer for MMI intradataset, but they used a different validation methodology from leave-one-subject-out, so the results may not be directly comparable.

4. CONCLUSION

In this paper we propose a method for generalization to person-independant and interdataset experiments for emotion recognition. The method improves classification rate over a baseline system by 15% when training on CK+ and testing on MMI. The method improves classification rate by 28% when training on MMI and testing on CK+. The method also achieves competitive intradataset results when compared to related work [1–5]. It was found that the surprise emotion causes the most change in facial appearance versus the other six basic emotions.

5. ACKNOWLEDGEMENTS

Support for this work was provided for in part by NSF grants 0727129, 0905671 and NSF IGERT: Video Bioinformatics grant DGE 0903667. The contents and information do not reflect the position or policy of the U.S. Government.

6. REFERENCES

[1] K. Ghanem, "Hidden markov models for modeling occurrence order of facial temporal dynamics," in *Conf. Adv. Concepts. Int'l. Vis. Sys.*, 2013.

[2] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simulataneous facial feature tracking and facial expression recognition," *IEEE Trans. IP*, vol. 22, no. 7, pp. 2559–2573, 2013.

[3] Y. Miao, R. Araujo, and M. S. Kamel, "Cross-domain facial expression recognition using supervised kernel mean matching," in *IEEE Int'l Conf. Machine Learning Applications*, 2012.

[4] A. Poursaberi, H. A. Noubari, M. Gavrilova, and S. N. Yanushkevich, "Gauss-Laguerre wavelet textural feature fusion with geometrical information for facial expression identification," *EURASIP J. Image and Video Processing*, vol. 17, 2013.

[5] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 920–992, 2012.

[6] A. J. O'Toole, F. Jiang, D. Roark, and H. Abdi, "Predicting human performance for face recognition," in *Face Processing: Advanced Models and methods*. 2006, pp. 293–317, Academic Press.

[7] M. Valstar, M. Mehu, J. Bihan, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," *IEEE Trans. SMC B*, vol. 42, no. 4, pp. 966 – 979, 2011.

[8] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Trans. Affective Computing*, 2014.

[9] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistic," *IEEE Trans. PAMI*, vol. 33, no. 10, pp. 1978–1990, 2011.

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

[11] P. Viola and M. Jones, "Robust real-time face detection," *Int'l. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[12] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 27, pp. 1–27, 2011.

[13] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: dense correspondence across scenes and its applications," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 978–994, 2011.

[14] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett, and J. R. Movellan, "Action unit recognition transfer across datasets," in *IEEE Conf. AFGR*, 2013.

[15] P. Lucey, J. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *IEEE Conf. CVPR*, 2010.

[16] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the MMI facial expression database," in *Corpora for Research on Emotion and Affect*, 2010.

[17] Y. Guo, G. Zhao, and M. Pietikainen, "Dynamic facial expression recognition using logitudinal facial expression atlases," in *IEEE Conf. ECCV*, 2012.

[18] A. R. Rivera, J. R. Castillo, and O. Chae, "Local directional number pattern for face anaylsis: face and expression recognition," *IEEE Trans. IP*, vol. 22, no. 5, pp. 1740–1752, 2013.

[19] B. Schuller, M. F. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012 Ű the continuous audio/visual emotion challenge," in *ACM Int'l. Conf. Multimodal Interaction*, 2012.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI*, vol. 24, no. 7, pp. 971–987, 2002.