Shape and Texture based Facial Action and Emotion Recognition (Demonstration)

Li Zhang, Kamlesh Mistry, Alamgir Hossain
Department of Computer Science and Digital Technologies
Northumbria University
Newcastle, NE1 8ST, UK
{li.zhang, kamlesh.mistry, alamgir.hossain}@northumbria.ac.uk

ABSTRACT

In this paper, we present an intelligent facial emotion recognition system with real-time face tracking for a humanoid robot. The system is able to detect facial actions and emotions from images with up to 60 degrees of pose variations. We employ the Active Appearance Model to perform real-time face tracking and extract both texture and geometric representations of images. A POSIT algorithm is also used to identify head rotations. The extracted texture and shape features are employed to detect 18 facial actions and seven basic emotions. The overall system is integrated with a humanoid robot platform to further extend its vision APIs. The system proved to be able to deal with challenging facial emotion recognition tasks with various pose variations.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *Perceptual reasoning*.

General Terms

Algorithms, Experimentation.

Keywords

Facial emotion recognition, Active Appearance Model, POSIT.

1. INTRODUCTION

It is envisaged that humanoid robots will play an increasingly active and engaged role in healthcare and educational settings for isolated elderly, autistic children and average users. However, how to conduct real-time efficient emotion detection from affective facial expressions, gestures and speech in a dynamic environment is still a challenging task for human robot interaction [1, 2, 3]. Especially, facial emotion perception has drawn strong attention in cognitive, neuroscience and computational intelligence fields. For example, Facial Action Coding System (FACS) [4] for measuring and describing facial behaviors has been developed by cognitive psychological scientists. It associated the momentary appearance changes with the action of muscles from the anatomical perspective. The system employed 46 Action Units (AU) to represent the muscular activities to describe and score facial expressions. Overall, FACS provided a versatile method to describe a wide range of facial behaviors, e.g. facial punctuators in conversation and emotional facial expressions.

This research aims to incorporate anatomical knowledge of FACS to guide facial emotion recognition. It extends the vision system of the latest humanoid robot, NAO NextGen H25 and employs physical cues embedded in both appearance and facial motions to detect AUs and emotions from frontal images with various pose variations. The Active Appearance Model (AAM) with POSIT (Pose from Orthography and Scaling with ITerations) is employed in this research in order to conduct efficient real-time face tracking and shape and texture feature extraction. Both a shape based neural network classifier and a texture based Local Binary Pattern algorithm are developed to detect the intensities of 18 facial actions respectively with the derived geometrical and textural features by AAM as inputs. The selected 18 AUs include AU1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 16, 17, 20, 23, 24, 25, and 26. We then recognize the following seven basic emotions using a neural network emotion classifier from the derived facial actions including happiness, sadness, surprise, fear, anger, contempt and disgust. The overall system is integrated with the latest robot C++ SDK under Ubuntu. Tested with images from the extended Cohn-Kanade (CK+) facial image database [5], the system achieves high recognition accuracies for AU and emotion detection with great robustness for pose variation detection.

2. AU AND EMOTION RECOGNITION

2.1 Shape and Texture Feature Extraction

In this research, in order to effectively capture discriminative physical cues embedded in both facial motion and texture deformation, we propose a shape and texture based AU and emotion recognition scheme with real-time face tracking for the NAO robot. This robot's original vision APIs are also able to perform basic object and face detection. However, it suffers from illumination changes and especially cannot deal with head rotations and textural feature extraction. This research work presented here is especially developed to further enhance the vision system of NAO. Since NAO's vision APIs allow the integration with the Open Computer Vision system (OpenCV), which contains various image processing algorithms, a Haar-based cascade classifier for face detection (also known as the Viola and Jones algorithm) provided by OpenCV is also employed in this research to perform real-time face tracking.

After a face is detected, an independent Active Appearance Model with the POSIT algorithm is used to extract geometrical and textural features from facial images. AAM was initially proposed

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

by Cootes et al. [6]. It is regarded as one of the most efficient model-based algorithms and proved to provide a better match to the texture in comparison to other models. AAM generates both shape and texture representations of deformable objects and has been widely used for face tracking, facial emotion recognition, and medical image segmentation. Moreover, an inspection of the images from the CK+ database shows that there are various minor head rotations from one frame to another. Therefore in order to perform more accurate face tracking and geometrical feature extraction, the POSIT algorithm is used to identify pose variations. POSIT was initially proposed by DeMenthon and Davis [7], which deals with finding the pose of an object from a single image. It uses an approximate pose to compute the scaled orthographic projections of feature points.

The CK+ database has overall 593 sequences of images with each image containing 68 2D landmarks. In order to build a robust and efficient AAM, we employ 2,926 images donated by 32 subjects extracted from the CK+ database for the training of AAM. These images represent transitional facial behaviours from neutral faces to peak emotional expressions. The training images and their corresponding 68 2D landmarks are both used as inputs to the training algorithm of AAM. Subsequently the generated AAM with an inverse compositional algorithm is used for AAM fitting. The task of AAM fitting is to search for the set of shape and appearance parameters which offer the best fitting between the trained model and the given test input image. Therefore the outputs of the fitting process are the bested fitted 68 2D shape landmarks and an extracted texture model of the input test image.

The 68 output landmarks by AAM are also further adjusted using POSIT by taking the pose of an object in the image into account. Finally these output shape and texture based representations of test images are respectively used as inputs to a geometric feature based neural network (NN) and the texture based Local Binary Pattern (LBP) algorithm to detect AUs.

2.2 Facial Action and Emotion Detection

As indicated above, both a geometrical feature based NN and the texture based LBP are used for AU detection. They are respectively trained with 327 images with AU and emotion annotations extracted from the CK+ database.

The shape based NN classifier uses the generated 68 landmarks from the fitting procedure as inputs and outputs the detected 18 AUs. Thus it has the 136 nodes in the input layer to accept the 68 2D landmarks, 18 nodes respectively in the hidden and output layers to represent the intensities of the detected 18 AUs. Moreover, LBP thresholds an *N* x *N* (e.g. N=3) neighbourhood of every single pixel to label the pixels of an image. The main motivation of using LBP is to describe only local features of an image instead of using the whole image as a high-dimensional vector. Thus the output texture vector by AAM is also used as inputs to LBP to detect AUs. In order to combine the shape and texture based classifications, the derived AUs from both of the approaches are ranked and seven AUs with higher intensities (i.e. stronger physical cues) are selected as the final outputs. After training with 327 images, both of the above shape and texture based classifiers are evaluated with the first 200 images taken from the database. The system achieved accuracies of 67.37%, 78.7%, and 85.73% respectively for the shape, texture and the combined approaches for AU detection. It shows that the combined approach with the consideration of both shape and texture information is able to offer more accurate AU detection with great robustness. Among the 18 AUs, majority of the AUs are well detected by this combined approach with AU10 (75%) and AU20 (60%) required further improvements.

A neural network emotion recognizer is also developed to detect emotions from the derived AUs for each test image. This NN emotion classifier has the following topology: 18 nodes in the input layer to accept intensities of 18 derived AUs, seven nodes respectively in the hidden and the output layers to indicate the selected seven emotions. The emotion classifier is also trained with 173 images with AU and emotion annotations. Using AUs derived from the above combined approach, the NN-based emotion classifier achieves 88.83% accuracy for the detection of the seven emotions respectively with ‘sadness’ (96.43%), ‘disgust’ (93.75%), ‘surprise’ (93.33%), ‘fear’ (93.33%), ‘anger’ (83.33%), ‘happiness’ (86.67%) and ‘contempt’ (75%).

The overall system is integrated with NAO’s latest C++ SDK, *Naoqi* SDK 1.14.3. It also allows local and remote cross tool-chain compilation with OpenCV 2.4.5 so that the compiled program is able to run both on desktops and the robot. The system also allows NAO to perform real-time facial emotion recognition for real testing subjects with various head rotations. NAO integrated with the system was also demonstrated in British Science Festival in 2013. In future work, clustering algorithms and other adaptive ensemble classifiers will also be employed to deal with compound and newly arrived novel unseen emotion class detection [8].

3. REFERENCES

[1] Zhang, L., Jiang, M., Farid, D., and Hossain, A.M. 2013. Intelligent Facial Emotion Recognition and Semantic-based Topic Detection for a Humanoid Robot. *Expert Systems with Applications*, Vol 40, Issue 13, 5160-5168.

[2] Zhang, L., Gillies, M., and Barnden, J.A. 2008. EMMA: an Automated Intelligent Actor in E-drama. In *Proceedings of International Conference on Intelligent User Interfaces*. Canary Islands, Spain. pp. 409-412.

[3] Zhang, L. 2013. Contextual and Active Learning-based Affect-sensing from Virtual Drama Improvisation. *ACM Transactions on Speech and Language Processing (TSLP)*, Vol 9, Issue 4, Article No. 8.

[4] Ekman, P., Friesen, W.V., and Hager, J.C. 2002. Facial Action Coding System, A Human Face.

[5] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. In *Proceedings of CVPR4HB*.

[6] Cootes, T.F., Edwards, G.J., and Taylor, C.J. 1999. Comparing Active Shape Models with Active Appearance Models. In *Proceedings of British Machine Vision Conference*. Vol. 1, 173-182.

[7] DeMenthon, D. and Davis, L.S. 1995. Model-Based Object Pose in 25 Lines of Code, *IJCV*, 15, 123-141.

[8] Farid, D., Zhang, L., Hossain, A.M., Rahman, C.M., Strachan, R., Sexton, G., and Dahal, K. 2013. An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams. *Expert Systems with Applications*, Vol 40, Issue 15. 5895-5906.