http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7305-7311, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7305.7311

Dynamic Feature Extraction for Facial Expression Recognition Based on Optical Flow

[1]Xibin Jia, [2]Shuangqiao Liu and [1,3]David M. W. Powers
[1]Beijing Municipal Key Laboratory for Multimedia and Intelligent Software Technology,
Beijing University of Technology,Beijing,100124, P.R. China
[2]College of Computer Science, Beijing University of Technology, Beijing 100124, P.R. China
[3]School of Computer Science, Engineering and Mathematics, Flinders University of South Australia,
Adelaide 5042, Australia

Abstract: Understanding facial expressions is a fundamental problem in affective computing that has the potential to impact both sides of a conversation with a computational agent. Currently static approaches based on techniques such as Gabor transformation represent the state of the art for identifying facial expressions but remain rather slow and unreliable. In this study we introduce and compare a dynamic technique based on optical flow versus this static Gabor baseline, as well as integrating information about static and dynamic facial characteristics into novel fused models. Rather than requiring complex Machine Learning, a simple and fast template model based on K-Means uses a Hidden Markov Model to provide context. The system is trained and evaluated on distinct subsets of the Cohn-Kanade database of adult faces which provides classifications into six basic expressions. Experimental results show that the dynamic expression feature extraction based on the optical flow for facial expression recognition has considerably improved recognition rate. Fusion of the optical flow dynamic expression features with Gabor static expression feature also increases recognition rate somewhat versus the static baseline but the hybrids tested are not competitive with the pure dynamic model.

Key words: Feature fusion, optical flow, local binary patterns, hidden Markov models, embodied conversational agents, affective computing

INTRODUCTION

Facial expressions are a form of nonverbal communication that convey semantic and affective information to an observer or interlocutor. They are a primary means of conveying social information between humans and thus facial recognition systems can also be expected to play a critical role in Human-Computer Interaction (HCI). In addition, facial expression understanding and synthesis should play an important part in the development interactive computer graphics systems based on human-like avatars and in promoting affective communication (Kana *et al*., 2006). Current research into facial expression recognition is primarily static based on single images (Wang *et al*., 2006) but in fact expressions and gestures are intrinsically dynamic which is to say they involve motion and change. Thus dynamic modeling of facial feature tracking points has the potential to considerably improve the accuracy of expression recognition and the naturalness of expression synthesis. Thus it is appropriate for computational facial expression recognition to exploit expression dynamics using optical flow algorithms.

In recent years, pattern recognition techniques based on optical flow has become an increasingly important topic in the field of pattern recognition, cognitive science, computer graphics and other disciplines of research. The earliest research of optical flow algorithm can be dating back to the 1950s. Wallach and Gibson (2001) proposed SFM (Structure From Motion) assumptions and Horn and Schunck (1981) developed the first □□truly effective optical flow calculation method, devising a theory of the development two-dimensional velocity field associated with the gray and introduction of optical flow constraint equation algorithms that has become a cornerstone of the optical flow algorithm. Optical flow technology is used for target tracking as an augmentation of machine vision recognition (Newman *et al*., 2010), in applications ranging from automatic vehicle navigation to weather monitoring. Compared with conventional techniques, the advantages of optical flow techniques are increased simplicity, timeliness and accuracy of target segmentation. Analysis of image motion thus plays a very important role in computer vision (Powers *et al*., 2008).

Standard approaches to Emotion Recognition include techniques based on PCA (Principal Components

Corresponding Author: Jia Xibin, College of Computer Science, #100 Ping Le Yuan, Chaoyang District, Beijing 100124, P.R. China Tel: +86 10 67392370; E-mail: jiaxibin@bjut.edu.cn.

Analysis) (Ali *et al*., 2011) and AAM (Active Appearance Model) (Cootes *et al*., 2002) models. The AAM models are trained on example images labeled with sets of landmarks to define the correspondences between images and across subjects (Cootes *et al*., 1998). One of the issues is that much of the work has concentrated on direct frontal imaging of the face, however Lanitis showed that a linear model was sufficient to simulate considerable change in viewpoint, as long as all the modeled features (the landmarks) remain visible (Lanitis *et al*., 1997).

In this study, we explore facial recognition using a standard dataset (Kanade *et al*., 2000). For that not only includes static images but video clips of computer users' facial expressions as they are told to act out six standard (basic) emotions. This allows exploiting the (dynamic) movement information from the videos using optical flow rather than just using (static) positional information derived from individual frames.

In this study we introduce a methodology for expression recognition in which optical flow calculations are performed across all frames of the video relating to the initiation, maintenance and release of the target expression. In the course of the analysis of this movement information, several expression features are identified to prepare for the recognition process. In this study we will introduce the optical flow algorithm used to extract and analyze the movement features, as well as subsequent processing steps including LBP (local binary patterns) to repair errors and a Hidden Markov Model, is used to provide state information for predicting the enacted emotion.

Furthermore, by combining these dynamic movement features with the static Gabor Wavelet Representation (Jia *et al*., 2013), two forms of feature fusion are shown to result in improvement over the static representation alone but currently they are not improving on the dynamic representation alone (Fig. 1).

EXTRACTING MOVEMENT FEATURES BASED ON THE OPTICAL FLOW ALGORITHM

Optical flow (Bradski and Kaebler, 2009), as is implied by its name, tracks changes in the optical properties of the scene that are functions not only of the object viewed but also of light, shadow, etc. Whereas different images and thus static features, may be strongly affected by such different illumination conditions, a sequence of images and thus dynamic features, tend to much less sensitive to such factors and it is thus relatively easy to identify the flow of interest points as objects, or parts of objects, move relative to the camera. In this work we are particularly concerned with the face and the articulatory organs including mouth and chin, cheekbones, eyes and eyebrows and we initialize specific characteristic points (features) using an AAM model, in order to allow tracking the motion of these key points associated with eyes, nose, mouth, etc. After marking of these characteristic points in an initial frame, the optical flow algorithm is applied to track the points, both detecting motion and calculate the relative movement information across the sequence of frames.

Choosing the characteristic points: As the characteristic points are the main parameters in the whole facial recognition process, the selection procedure is crucial and although an unsupervised approach may be used to discover points of interest (Lang *et al*., 2013). The system must ensure that these selected characteristic points are representative for human faces' expressions and, at the same time, not too susceptible to luminance changes. As such, the system applies the AAM method together with manual adjustments to select the characteristic points to be used in the optical flow calculation. (Fig. 2).

Lucas-Kanade method of optical flow: The Lucas-Kanade optical flow method (LK method) (Bradski and Kaebler, 2009) is a widely used differential method for optical flow estimation in computer vision. It assumes that the flow is essentially constant in a local neighborhood of the pixel under consideration, under three assumed conditions: Constant luminance, small movements and no change in the background. Based on these three basic assumptions, the LK method solves the basic optical flow equations for all the pixels in that neighborhood, by the least squares criterion. OpenCV (Bradski and Kaebler, 2009) provides two built-in functions associated with the LK method. The

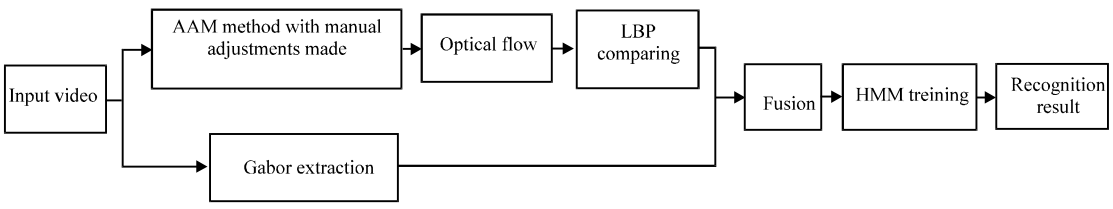

Fig. 1: Framework of the proposed facial expression recognition methods

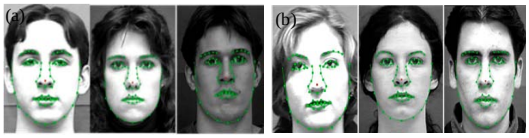

Fig. 2: Sample characteristic points selected using the AAM method with manual adjustments

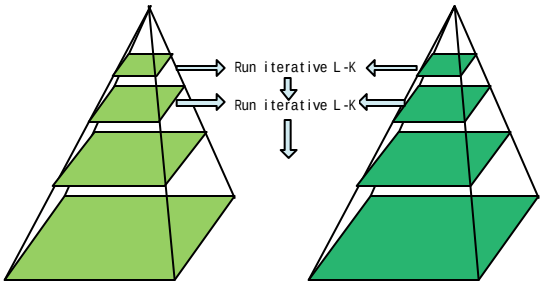

Fig. 3: Pyramidal implementation of LK method (Bradski and Kaebler, 2009)

first calculates the basic optical flow equations on all pixels both vertically and horizontally which further gives information on the flow of the whole image. The other bases on a pyramidal and iterative model to process a group of characteristic points in the image by combining information from several nearby pixels which narrows down these characteristic points that do not meet the three conditions and creates a faster and more stable way to keep track of the flow. Based on the analysis of the built-in functions, the system ultimately uses the pyramid Lucas-Kanade algorithm to calculate the optical flow equations in the facial recognition process.

Pyramidal implementation of LK method: An efficient approach to implementing the LK method is to build a pyramid representation (Bradski and Kaebler, 2009). For each initial feature point represented in the topmost layer of the image pyramid model (layer k) we find a matching point and then use this point from the layer k as the initial estimate of displacement in the next layer of the image pyramid (layer k-1) as we search for matching points iterating until we reach layer 0 (original image) of the image pyramid model (Fig. 3).

The pyramid principle is based on solving the partial derivative of displacement which calculates the displacement vector of the center of each layer of the pyramid, by tracking trends about the features' motion.

In the function cvCalcOpticalFlowPyrLK, the velocity and direction of the feature point is characterized by an equivalent displacement and direction, as shown in the following Eq. 1:

$$E(x) = E(d_x, d_y) = \sum_{x=ux+wx}^{ux+wx} \sum_{y=uy+wy}^{uy+wy} (I(x, y) - J(x + d_x, y + dy))^2 \quad (1)$$

In Eq. 1, E is the magnitude of the error, w_x and w_y are the dimensions of the window, is a pixel in the previous frame image, is a pixel in the current frame image. find represents the partial derivative which we constrain to 0, so as to obtain the corresponding d.,, when d is 0, McLaughlin, the equation (1) can be reduced to Eq. 2:

$$G = \sum_{x=ux+wx}^{ux+wx} \sum_{x=ux-wx}^{ux+wx} \begin{bmatrix} I_x^2 & I^{L_x L_y} \\ I_x I_y & I_y 2 \end{bmatrix}$$
$$b = \sum_{x=ux-wx}^{ux+wx} \sum_{x=ux-wx}^{ux+wx} \begin{bmatrix} I_x & I_t \\ I_y & I_t \end{bmatrix} \quad (2)$$

is the optimal solution and the solution process requires an iterative processing.

IMPROVED OPTICAL FLOW TRACKING ALGORITHM BY LOCAL TEXTURE SIMILARITY

One of the assumptions of the optical flow algorithm concerns constant brightness but in practice illumination changes and shadows occur, including self-cast shadows by parts of the face or the articulators moving into or out of shadow. The result is that optical flow can't always track the movement of the characteristic points accurately,. Thus significant errors can occur. To address this issue, this study introduces an improved optical flow algorithm using a window around the extracted feature points in which local textural similarity constraints are checked. That is within a fixed area around the source and destination points, the feature points are verified according the detailed texture.

Improved optical flow: To reduce the effects of any unforeseeable or temporary luminance changes while tracking the characteristic points, we introduce a test of local texture similarity. This helps replace the wrong pixel tracked by the LK method using the texture similarity test. Local binary patterns: We use Local Binary Patterns (LBP) to calculate the texture classification before and after the estimated movement (Fig. 4). In a 3-by-3 square region, the grey level of the central pixel is set as the threshold (which in this case is 48) and is compared with the grey levels of the adjacent eight pixels. A higher grey level is marked 1 or 0 otherwise. The local binary pattern is then created through a traversal of the whole image.

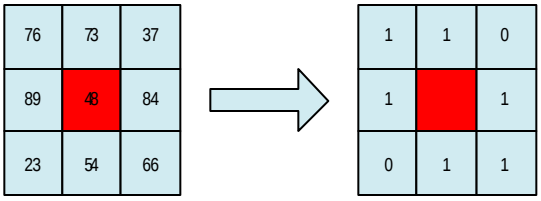

Fig. 4: 3-by-3 window in local binary patterns

Using the LBP, the type feature is used as an important input in tracking the characteristic points. A 5-by-5 square region is considered near each characteristic point. The texture classification calculations are done before and after the flow. A counter is also created for further process. If the matching pixels have the same LBP feature, the counter is add by 1. Hence the counter ranges from a minimum of 0 to a maximum of 24 and is used as an important factor in flow tracking.

This algorithm helps improve the original pyramid optical flow method by comparing the texture similarity. Judging by users' experience, the threshold of the counter is set at 18 in this study. Moreover, the square of the distance between the flow must be less than 10 to be considered as a match. If these two conditions are not met, the characteristic points will be replaced by another pair.

Empirically, we find typically two or three of the 70 reference points are inaccurately tracked *per frame*. The LBP technique manages to correct about more than 29% of these errors than without it.

DESIGNING AND REALIZING OF THE TECHNOLOGY OF FEATURE FUSION

Design: To verify the utility of recognition of facial expressions using dynamic information and to explore whether the dynamic technology is complementary to and usable with the static information, this study explores two methods of integrating the dynamic and static facial feature information. The first method involves simple concatenation or 'stitching' of the dynamic and static feature information and the other is using the technique of semantic alignment (Mitchell, 2010).

Simple fusion: For feature information fusion it is important to take into account the different characteristics of static and dynamic feature. Using the Euclidean distance as a measure, concatenation of different kinds of information on different scales will lead a serious biases that can reduce classification accuracy. Therefore, we propose a joint static and dynamic feature normalization process, before concatenating them into training samples, so that each component of the feature vector has the same internal position in the similarity metric.

For fusion of the dynamic and static features, we normalize the respective information prior to stitching. The normalization is designed to ensure the dynamic and static feature vectors have comparable significance within the similarity or distance measures and we use a linear normalization by the largest value of each attribute. Formally, given a sample with feature vector of dynamic information samples F_D = {f1, f2.....fm}; feature vector of static information samples F_G {f1, f2.....fm} value of each attribute is divided by the maximum value of the corresponding attribute, joining the normalized vectors together to form a new joint samples:

$$F_{D\text{-}G} = \left\{ \frac{F_D}{F_D, \max}, \frac{F_G}{F_G, \max} \right\}$$

This process of normalization is now followed by the K-means clustering and HMM training steps (Fig. 5).

Semantic fusion: Semantic refers to the meaningful description or explanation of things in the real world and semantic alignment refers to matching of description of the same object or phenomenon as recognized from different input data measures. The reason for performing semantic alignment is that different inputs can only be fused together if the inputs refer to the same object or phenomena. However, when the sensors are of different types, the observation may refer to different phenomena. If this is true, then it is not possible to perform data fusion unless the input data are semantically aligned to a common object or phenomena. Therefore, we propose to find the same expression feature of different types of data (different sensors) described and combine them to form a sample of fused features, so as to achieve standards of the semantic alignment.

A related process is prototypical alignment, where a specific attribute vector is replaced by a more typical prototype, e.g. the centroid of a cluster.

To implement this technique of semantic alignment or prototype and facilitate the facial dynamic and static features information fusing together, we separately cluster the two kinds of features and calculate the centroid of each cluster as a prototype. For prototype alignment we use the centroids spliced together rather than the original vectors, as the input to subsequent stages in processing. For prototype alignment an unsupervised process is meaningful and we use k-means classification where k= 6 corresponding to the target number of emotional states

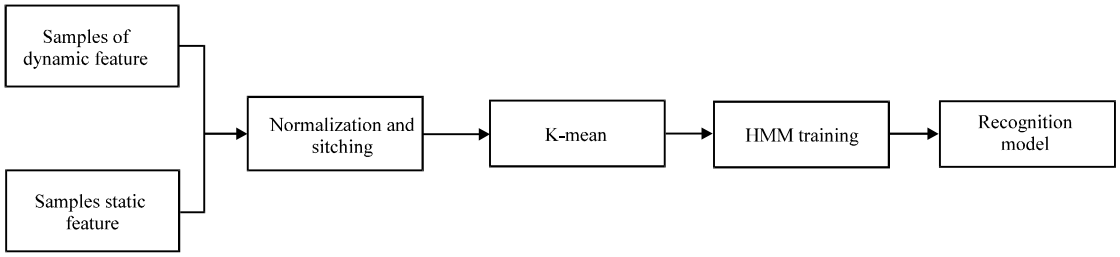

Fig. 5: processing of dynamic and static feature fusion with normalized stitching

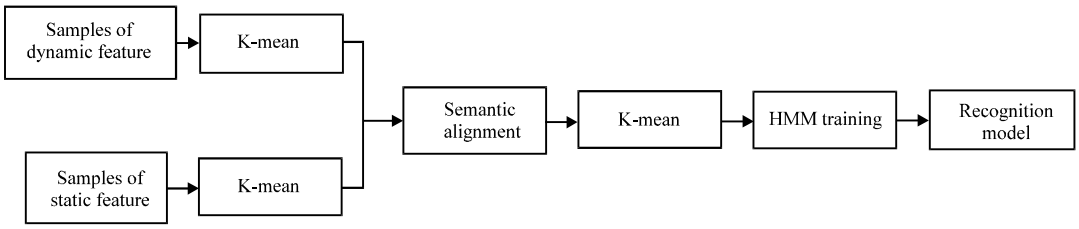

Fig. 6: processing of dynamic and static feature fusion with semantic alignment

and alternative dynamic implementation is based on this method of face recognition (Fig. 6). An alternative approach would be to use a supervised classification process such as k-NN and use these purer templates for stitching into a fused attribute vector.

A strict semantic alignment approach would require reconciling the different classifications associated with the dynamic and static features in a way that minimized error and investigation of such an approach is left for future work. The current approach is however more powerful as it has the potential of recognizing that an emotional state combines two (or more) emotions and/or involves a transition between emotions. In addition, the unsupervised approach allows having k>K, that is it is possible to produce more clusters than the ostensible number of basic emotions and thus distinguish different shades of variants of emotion. These considerations also provide rich scope for further exploration in future work.

EXPERIMENTS AND DISCUSSION

Four distinct systems were trained using 40 training images from the Cohn-Kanade database: Gabor baseline, optical flow only, early fusion and clustered fusion. Experiment 1 (Fig.7a) shows the result yielded by the original pyramid optical flow method. Experiment 2 (Fig.7b) shows the result with the improvement of texture similarity matching. Notice that these two calculations are captured at the same frame but experiment 2 shows a dramatic improvement over experiment 1.

Fig. 7(a-b): (a) Experiment1 (b) Experiment 2

For the two methods of fusion, the result of two kinds of joint feature, dynamic feature and static feature, the

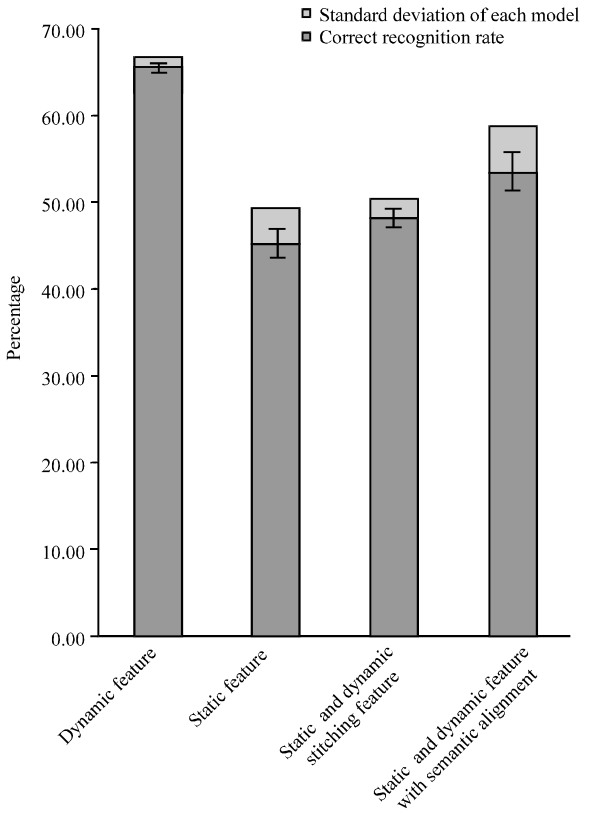

Fig. 8: Histogram of four face recognition results with feature model, showing standard deviations and errors bars of 1 standard error

system takes four kinds of contrastive experiments respectively based on different feature of training model, like experiment (Fig.8). The total sample number is 55 and they are divided by six kinds of expressions.

The experimental results show that the expression of dynamic feature for facial expression recognition effect is the best. Meanwhile the results prove that dynamic feature for the recognition and analysis of facial expression has the decisive effect. For the two methods of integrating with the dynamic feature and static feature on characteristic level, the technology of stitching the facial dynamic and static feature information fusion has the higher recognition rate but is still not competitive with dynamic features alone implying that there is further work needed to achieve effective fusion.

For a more detailed description of the facial expression recognition based on four kinds of facial expression feature, the study implements the comparative experiments based on four feature models to identify the effect of the expression of each basic expression (Fig. 9). In this pilot experiment we selected examples with good lighting and low variation in illumination with typically 4

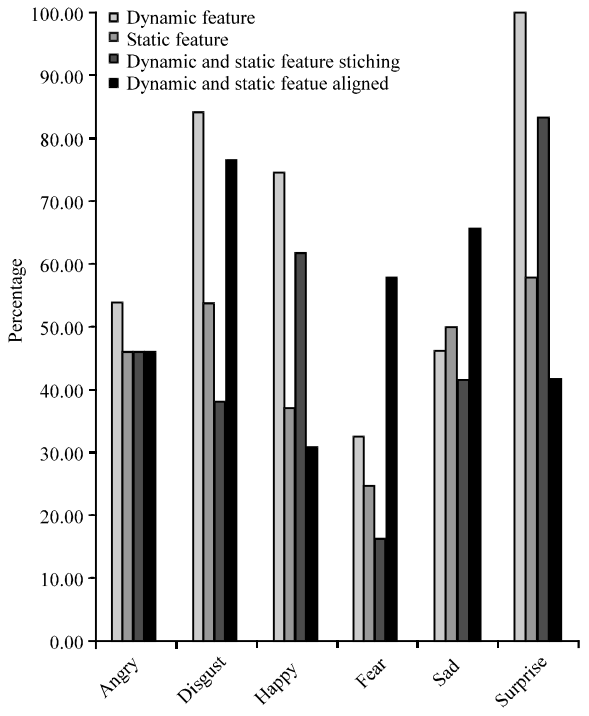

Fig. 9: Correct recognition rate based in four feature model

test samples and 12 training samples per emotion (but for technical reasons relating to the preprocessing required to deal with illumination issues, happiness had 16 training samples and surprise only 3 test samples).

The result of experiments shows that the correct identification rate of dynamic feature recognition (blue histogram) is the highest, especially the identification of a surprised expression, the correct rate can reach 100%. Looking at the identify situations of longitudinal comparison of six basic facial expressions, expression of surprise has the higher correct rate of recognition than others in the same way. On the contrary, the expression of fear which is much complex has the lower recognition rate. Looking at the identify situations of horizontal comparison, each method has its own good at recognition.

CONCLUSION

In this study, the facial expression of dynamic feature extraction is studied and implemented based on optical flow. At the same time, on the basis of the static image processing and the technology of information fusion, making the combination of the dynamic feature extracted by optical flow and the static feature extracted by Gabor, treating the Hidden Markov as classifier, training four kinds of feature models, that is dynamic feature model,

static feature model and two kinds of fusion models. Finally, identifying the facial expression with testing video and contrasting the recognition effect based on different feature models. Through the analysis of experimental results, this study proves that trend of facial movement plays important role in facial expression recognition and establishes an effective dynamic feature can improve the recognition results. At the same time, the result also shows that dynamic and static fused features especially based on proposed semantic alignment has higher recognition rate in some expressions viz. fear and sad which is normally hard to be identified.

In the further job, we will use more public dataset to test our proposed method. The approach becomes more efficient if the method can be adaptive and personalized based on Optical Flow. The performance of facial expression recognition with our proposed method on the different condition will be analyzed, like illumination, background, skin color, etc. The components displaying better performance and possessing complementary information in contributing for the facial expression recognition will be discussed. Fusion method with different kinds of feature on the above components will be researched to improve the facial expression recognition results.

ACKNOWLEDGMENTS

This work is supported by the Chinese Natural Science Foundation under Grants No.61070117, the Beijing Natural Science Foundation under Grant No.4122004, Specialized Research Fund for the Doctoral Program of Higher Education (20121103110031), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions.

REFERENCES

Ali, H.B., D.M.W. Powers and R. Leibbrandt and T. Lewis, 2011. Comparison of Region Based and Weighted Principal Component Analysis and Locally Salient ICA in Terms of Facial Expression Recognition. In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Lee, R. (Ed.). Springer-Verlag, Berlin, Heidelberg, pp: 81-89.

Bradski, G. and A. Kaebler, 2009. Open CV Learning. 3rd Edn., Tsinghua University Press, Beijing, China, pp: 350-355.

Cootes, T.F., G.J. Edwards and C.J. Taylor, 1998. Active appearance models. Proceedings of the 5th European Conference on Computer Vision, June 2-6, 1998, Freiburg, Germany, pp: 484-498.

Cootes, T.F., G.V. Wheeler, K.N. Walker and C.J. Taylor, 2002. View-based active appearance models. Image Vision Comput., 20: 657-664.

Horn, B.K.P. and B.G. Schunck, 1981. Determining optical flow. Artif. Intell., 17: 185-203.

Jia, X.B., X.Y. Bao, D.M.W. Powers and Y.J. Li, 2013. Facial expression recognition based on block Gabor wavelet fusion feature. J. Converg. Inform. Technol., 8: 282-289.

Kana, R.K., T.A. Keller, V.L. Cherkassky, N.J. Minshew and M.A. Just, 2006. Sentence comprehension in autism: Thinking in pictures with decreased functional connectivity. Brain, 129: 2484-2493.

Kanade, T., J.F. Cohn and Y. Tian, 2000. Comprehensive database for facial expression analysis. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, March 26-30, 2000, Grenoble, France, pp: 46-53.

Lang, S.R., M.H. Luerssen and D.M.W. Powers, 2013. Automated evaluation of interest point detectors. Proceedings of the IEEE/ACIS 12th International Conference on Computer and Information Science, June 16-20, 2013, Niigata, Japan, pp: 443-447.

Lanitis, C.J., T.F. Taylor and T.F. Coates, 1997. Automatic interpretation and coding of face images using flexible models. IEEE Trans. Pattern Anal. Machine Intell., 19: 743-756.

Mitchell, H.B., 2010. Data Fusion: Concepts and Ideas. Springer, New York, USA., pp: 126-128.

Newman, W., D. Franzel, T. Matsumoto, R. Leibbrandt, T. Lewis, M. Luerssen and D.M.W. Powers, 2010. Hybrid world object tracking for a virtual teaching agent. Proceedings of the IEEE International Joint Conference on Neural Networks, July 18-30, 2010, Barcelona, Spain, pp: 2244-2252.

Powers, D.M.W., R.E. Leibbrandt, D. Pfitzner, M.H. Luerssen, T.W. Lewis, A. Abrahamyan and K. Stevens, 2008. Language teaching in a mixed reality games environment. Proceedings of the 1st International Conference on Pervasive Technologies Related to Assistive Environments, July 15-19, 2008, Athens, Greece.

Wallach, J.C. and L.G. Gibson, 2001. Mechanical behavior of a three-dimensional truss material. Int. J. Solids Struct., 3: 7181-7196.

Wang, J., L.J. Yin, X.Z. Wei and Y. Sun, 2006. 3D facial expression recognition based on primitive surface feature distribution. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 17-22, 2006, New York, USA., pp: 1399-1406.