International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 12, December 2015

Synthesizing Face Video of a Target Subject

[1]Preetam S. Varude, [2]Sujata S. Agrawal
[1]preetam.varude@gmail.com,[2]sujata.agrawal@rediffmail.com

Abstract:- **Synthesizing a face video of a target subject that nothing but the mimicry of the expressions of a source subject in the input video is facial expression retargeting in video. Facial expression retargeting has applications areas like dummy pictures. In retargeting problem, it uses the facial expression video of one subject as input to synthesize new facial expressions of another subject, therefore it is more reasonable to include the facial expression of different subjects in the training and test datasets. That is, two datasets should not contain the expressions of the same subject for the application. This paper includes required results of different facial expression retargeting techniques.**

Keywords : **Facial expression, expression retargeting, expression synthesis, expression similarity, tensor factorization..**

I.INTRODUCTION

Synthesizing facial expressions of a target subject that exhibit the same expressions of a source subject is referred as facial expression retargeting or performance-driven facial animation. To be a successful retargeting system, it should meet following three criteria: (1) *similarity*, meaning that the synthesized expressions should be perceptually close to those in the input performance, although the subjects are different; (2) *naturalness*, meaning that the synthesized expressions should look natural without noticeable artifacts; and (3) *efficiency*, the proposed system should require minimal user input and is general enough to handle various subjects [base paper].

These systems have drawn plenty of attention since the 1980s, yet it still not so precise. Methods devised previously often fail to meet all requirements mentioned above simultaneously. Many previous approaches such as [1],[2] focus on the similarity criterion, but they require too much user interaction for generating the output expressions. On the other hand, methods like performance-based facial animation [3] focus on photo-realistic rendering of the synthesized expressions, but they require accurate 3D face models of the subjects, which are hard to obtain and require special devices and setups.

Recently, data-driven approaches have shown great potential in various synthesis problems such as creating human motions [4] and completing occluded faces [5]. Inspired from this methodology, one pre-captured video database of the target subject is used to achieve photo-realistic expression retargeting. A database includes some basic expressions such as neutral, angry, disgust, fear, happiness, sadness, and surprise. Since video frames in the database contain the ground-truth appearance of the target person under various expressions, they can be used as strong appearance priors for rendering new expressions. This allows developing an efficient facial expression retargeting system without using accurate 3D models which are hard to obtain.

The rest of the paper is organized as follows: In next Section II, different techniques used for real facial 3D models are discussed. Section III describes the various methods to be considered for synthesizing the face video.

In order to separate expression from identity changes, Yang et al. proposed a method to jointly fit a pair of face images from the same person. However, their method assumes a single dominant expression for each pair whereas our method can handle a general linear mixture of expressions and identities. We achieve that using a 3 mode tensor model that relates expression, identity and the location of the tracked feature points. A few related tensor models were introduced in the past. Vasilescu and Terzopoulos proposed tensor face to model the variations in frontal face images. Their model was used for face recognition and achieved better accuracy than PCA. Vlasic et al. built a 3D tensor model for face animation that related expressions, identity and poses.

However, these methods do not show how to directly solve the model coefficients for a new person, not in the dataset. In addition, they were not designed to work with general video sequences whereas we explicitly solve for a single identity for the entire video and require smooth variations of expression and pose for a more robust and realistic solution. Dale *et al.* extended Vlasic's approach for replace facial performance in video. They could transfer expressions to a different subject that is not from the training set. However, their system requires accurate initialization of the identity parameters that relies on commercial face reconstruction software, as well as on user interaction in one or more key frames. To set the identity they use just the first frame, while our method is more robust to noise as we infer the identity by jointly fitting all frames of the video.

Fig: 1 Facial features

In fig.1, the green curves connect all AAM features and the pink curve is the contour of the projected face geometry. The short red lines show the landmarks projected onto the face contour.

II. OVERVIEW

A. Constrained Local Model

A real time facial puppetry system is presented and when compared with existing systems, no special hardware is required, works in real time (23 frames-per-second), and requires only a single image of the avatar and user. A real-time 3D non-rigid tracking system is used for capturing user's facial expression. Combining a generic expression model with synthetically generated examples provide expression transfer that better capture person specific characteristics. Performance evaluation of system is based on avatars of real people as well as masks and cartoon characters.

B. 3D Multilinear Model

2D morphable model based automatic face replacement in video is used in this method. This approach contains three important modules: face alignment, face morph, and face fusion. The Active Shape Models (ASM) is adopted to source image and target frames for face alignment in given a source image and target video. Then with the help of a 2D morphable model, the source face shape is warped to match the target face shape. The color and lighting adjustments of source face are done to keep consistent with those of target face, and flawlessly merged in the target face. This approach is fully automatic i.e. without user interference, and provides natural and realistic results.

III. PROPOSED WORK

The system block diagram is shown in Fig. 2. Given the input sequence of the query subject, and the expression database of the target subject, we first identify the neutral expression for both subjects, which will be used in the expression metric for retrieval. The expression metric measures not only the motion from the neutral frame to an expression frame, but also the temporal motion velocity at the expression frame. The metric is further improved by a learning-based approach. Using the metric the system produces a retrieved sequence. The system also generates a synthesized sequence using expression mapping, and finally combine these two sequences together to produce the final result.

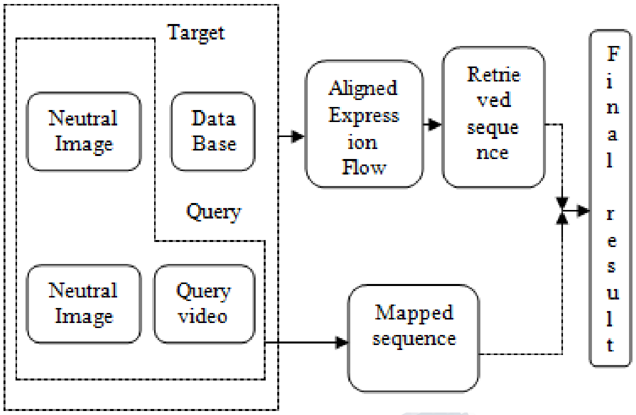

Fig. 2: Block diagram for proposed method

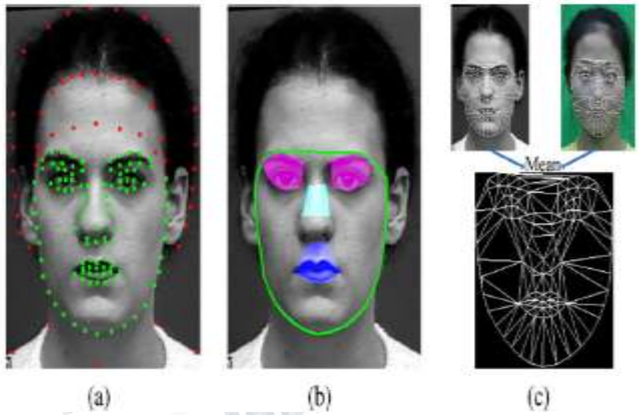

Fig 3. Initial processing of neutral face

A. Optical flow-based descriptor

To accurately align the optical flow fields, we first extract facial landmarks using the active shape model (asm) (fig. 3(a)), then remove the non-deformation offset such as 2d translation, rotation, and scaling, by computing a similarity transform matrix between two models on the nose region (fig. 3(b)), which is mostly invariant to expression changes. To account for the difference between two facial shapes, we build a piece-wise affine mapping function via delaunay triangulation, as shown in fig. 3(c). Using this mapping function, we map both flow fields to a common reference face shape , which is a pre-defined canonical shape. If such a canonical shape is not available, we use the mean shape of the neutral faces of both the query and the target subject as , as shown in fig. 3(c).

fig 3(a) shows green markers are automatically detected, and red ones are manually labeled used for expression mapping, 3(b) shows the green contour shows the face region. The eye, nose and mouth regions are marked in magenta, cyan, and blue, respectively. 3(c) shows the mean shape of query neutral face and target neutral face can be regarded as a reference shape.

B. *Incorporating expression velocity*

The distance function only considers the static expression distance. When dealing with video frames, the velocity of expression changes at each moment also needs to be taken into account. In other words, when comparing a query frame and a database frame, we expect not only the static expression distance between them is minimized, but also the expression change momentum at these two times needs to be matched. This will greatly improve the temporal coherence of the retrieved frames.

C. *Expression mapping*

Using this method, given a neutral image x and an expression image x' of the same subject as example data, the expression, including its associated appearance details (e.g., winkles around mouth corners when smiling), can be transferred to the neutral face y of another subject to create a new expression image y' . The key idea of eri is to utilize illumination changes to describe facial expression changes. It then proposed a new method called expression mapping image (emi) to synthesize novel facial expressions of the target subjectgiven the examples of the query subject. However, the emi sequence just consists of independently-generated emi images.

IV. SIMULATION RESULTS

To evaluate the system as a whole, we collect the databases of three target subjects. The videos of two of the target subjects, one male and one female, denoted as and, respectively, are collected by ourselves. These two non-professionals are asked to perform some basic expressions for one minute. Thus each database video contains roughly 1500 frames. The third target subject is a female subject from the ck+ dataset [36], whose database consists of 11 short sequences (220 frames). Also collected facial performance videos of other query subjects as input data to the system.

TABLE I: SUBJECTIVE EVALUATION RESULTS

Method	T1	T2	S130
[7]	1.26	01.58	1.41
ST-EMI	2.85	1.91	3.45
Proposed Method	4.02	4.52	4.06

The result of target subject and driven by a common video of a male subject. The result of target subject driven by a female subject. From the results we can see that, first, our system is able to synthesize new expressions that are not performed in the database, such as. This conclusion is also supported by the result with a small target database. Secondly, our system can handle expressions performed at different speeds.

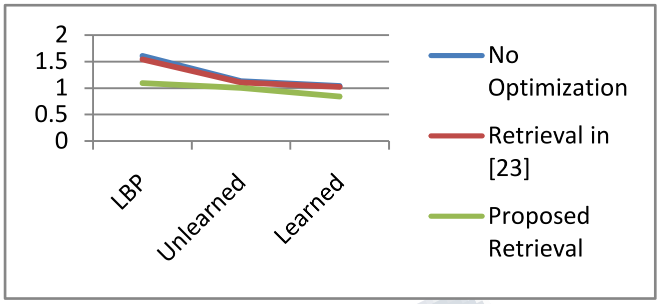

Fig. 4: Objective evaluation results

Optimization-based retrieval strategy outperforms the retrieval strategy proposed in [7]. In particular, with our retrieval method, the temporal coherence of the LBP feature is improved by 31.9%.

V.CONCLUSION

A learning-based expression similarity metric to measure facial expression similarity between different subjects.

It then propose the optimization based approach for generating a retrieved sequence which matches with the expression performance of the input video. system significantly outperforms previous approaches on achieving realistic, temporally coherent and accurate expression transfer results.

REFERENCES

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc.IEEE Workshop CVPR for Human Communicative Behavior Analysis*, 2010, pp. 94–101.

[2] F. Xu, Y. Liu, C. Stoll, J. Tompkin, G. Bharaj, Q. Dai, H.-P. Seidel, J. Kautz, and C. Theobalt, "Video-based characters: Creating new human performances from a multi-view video database," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 32:1–32:10, 2011.

[3] J. M. Saragih, S. Lucey, and J. F. Cohn, "Real-time avatar animation from a single image," in *Proc. IEEE Int. Conf. Automatic Face andGesture Recognition*, 2011, pp. 117–124.

[4] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, and H. Pfister, "Video face replacement," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 130:1–130:10, 2011.

[5] T. Weise, S. Bouaziz, H. Li, and M. Pauly, “Realtime performance- based facial animation,” *ACM Trans. Graph.*, vol. 30, no. 4, pp. 77:1–77:10, 2011.

[6] Y. Seol, J. Lewis, J. Seo, B. Choi, K. Anjyo, and J. Noh, “Spacetime expression cloning for blendshapes,” *ACM Trans. Graph.*, vol. 31, no. 2, pp. 14:1–14:12, 2012.

[7] F. Yang, L. Bourdev, E. Shechtman, J. Wang, and D. Metaxas, “Facial expression editing in video using a temporally-smooth factorization,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 861–868.

[8] K. Li, F.Xu, J. Wang,Q.Dai, andY. Liu, “A data-driven approach for facial expression synthesis in video,” in *Proc. IEEE Conf. ComputerVision and Pattern Recognition*, 2012, pp. 57–64.

.

● ● ●