iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** III **Month of publication:** March 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Synthesis for a Face Video of Target Subject

Preetam S. Varude[1], S. S. Agrawal[2]
Department of E&TC University of Pune,India

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 survey 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. Finally, Section IV concludes the paper.

II. RELATED WORK

Early research on facial animation heavily depends on user specified facial features. Litwinowicz and Williams [1] animated the image with line drawings by texture mapping. Beier and Neely [2] presented a metamorphosis technique by providing line pairs on face image. Liu *et al.* [6] proposed the concept of expression ratio image, i.e., the illumination changes from the neutral image to an expression image, which is applied to another person's neutral face to achieve expression transfer purpose. Zhang *et al.* [8] generated the photorealistic expressions through a combination of examples images in each face sub region. The synthesized texture is inferred by applying the geometry relationship in each sub region to the texture of example images.

A. Constrained Local Model

Training images with labelled landmarks to capture the facial appearance statistics is used in Constrained Local Model (CLM) [3].

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

By using a 3D multilinear model, Dale presented a system which replaces the faces in videos that are of approximately the same appearances, expressions, and poses [4].

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.

C. Spatio-temporal Multilinear Model.

The problem of editing facial expression in video can be addressed as overstating, tempering or substituting the expression with a different one in some parts of the video. To achieve this, a tensor-based 3D face geometry reconstruction method, which fits a 3D model for each video frame is developed, with the constraint that all models have the same identity and requiring temporal continuity of pose and expression. The differences between the underlying 3D shapes capture only changes in expression and pose with the identity constraint. Various expression editing tasks in video can be achieved by combining face reordering with face warping, where the warp is induced by projecting differences in 3D face shapes into the image plane. Analogously, the identity can be used while setting expression and pose. Method does effectively edit the expressions and identity in video in a temporally-coherent way with high fidelity.

Video on the web is growing at astonishing rates. In 2011, on YouTube alone, every minute people upload 8 years of video content. While video capture, bandwidth and storage have become easier over time, semantic editing of video content remains a very challenging problem. Consider the problem of making a person smile in an image. This is not as simple as cutting a smile from another image, pasting it and blending the results. The entire face changes its shape, the chin becomes wider and the eyes become narrower. The appearance depends on the pose of the person as well as the identity: everyone smiles in a unique way. While a Photoshop expert with sufficient amount of time could change one’s expression in an image, doing so in video is prohibitively expensive. The time dimension adds new sets of constraints, such as temporal coherence, and the temporal “signature” of an expression.

Fig.1. Magnify or suppress an expression in video.

Fig.1 shows magnify or suppress an expression in video. Middle Frames are from the original video. Top frames are synthesized frames in which the smile is suppressed. Bottom frames are synthesized frames in which the smile is magnified.

The main goal is to allow for semantic-level editing of expressions in video, such as magnifying a smile (Fig. 2.1) or an expression of fear, inserting an expression, or replacing unwanted expressions, such as an eye roll or facial tics. In addition, we can change the facial structure of the person, such as widen the chin or narrow the forehead, while preserving the pose and expression. We propose a new face fitting algorithm which takes a video of a person's face and decomposes it into identity, pose and expression. This decomposition allows us to make high-level edits to the video by changing these parameters and synthesizing a new video. We define our task as an energy minimization problem with the constraints of temporal coherence of the pose and expression and unique identity of the person in all frames. We model the face geometry over time using 3-mode tensor model, which can only deform in low-dimensional tensor space. Our method results in high fidelity reconstruction and has some robustness to view point variation.

III. SYNTHESIZING METHODS

A. *Tensor factorization methods for faces*

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.

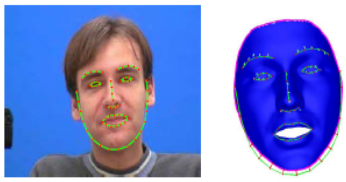


Fig. 2. Left: Facial features detected and tracked by Active Appearance Model (AAM), Right: Updating face contour landmark correspondences.

In fig.2, 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.

B. *Dynamic time warping*

We treat the input video as a dataset with expressions and apply the DTW method to map the sequence of new to the dataset. The distance map is computed as the Euclidean distance in the expression subspace. In the original video the subject changes expression from neutral to full smile. The original expression coefficients are scaled to neutralize the smile. The new sequence, with expression only maps the first half of the original video. Therefore, the result video only shows a half smile.

C. *Learning based expression metric*

In this section, we describe our expression similarity metric that takes into account the appearance difference between different subjects. Our metric is built upon an optical flow based descriptor, which can capture subtle facial expression changes between two facial images. We further show that for video frames, good expression matching requires not only their static facial

expression distance, but also higher order statistics such as the velocity of expression changes at individual frames. By incorporating expression velocity, we construct an Euclidean facial expression similarity metric. Finally, we present a metric learning approach which improves the accuracy and robustness of the proposed metric.

D. Optical Flow-Based Descriptor

Given an expression image of the query subject and another expression image of the target subject, our goal is to compute the expression distance (or similarity) between expression image of query and target subject. We take advantage of the corresponding neutral face of person and of to achieve this goal, which are manually selected from the videos. This needs to be done only once for each subject. The motion filed between neutral and expression image of query subject and (similarly between these) can well capture the facial difference caused by the expression shown in expression image of query subject. We thus estimate the motion using an existing optical flow approach and build the metric upon it.

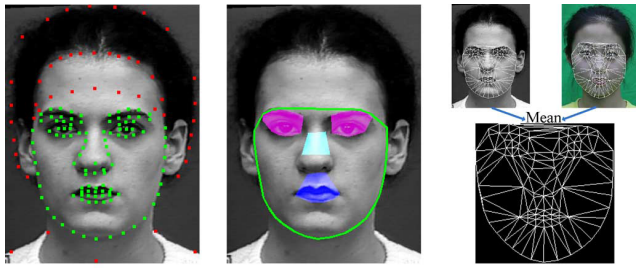


Fig. 3. Initial processing of the neutral face.

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

E. Incorporating Expression Velocity

The distance function 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, it expects 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.

F. Metric Learning

The expression metric proposed in (8) is a Euclidean distance. However, the flow-based metric does not encode expression semantics, thus may not be consistent with the human perception of facial expressions. For instance, consider the case that one subject changes his expression from sad to a subtle smile, and then to a big smile. The distance from sad to a subtle smile may be the same as the distance from a subtle smile to a big one, according to the flow-based metric, however the former is usually considered as a bigger mode change for human perception. Under this observation, we propose to use a learning-based approach to make the expression metric be more consistent with human perception.

IV. COMPARATIVE ANALYSIS

In this section, a comparison of described techniques is discussed to verify the effectiveness of each technique of synthesizing a video of target subject. CLM needs only a single image of the avatar and user. A real-time 3D non-rigid tracking system is used for capturing the user's facial expression. Performance evaluation of the system is done based on avatars of real people and also on masks and cartoon characters. With a 2D morphable model, in 3D multilinear model, source face shape is warped to match the target face shape. It is fully automatic i.e. without user interference, and gives natural and realistic results. Spatio temporal model results in high fidelity reconstruction and has some robustness to view point variation. The tensor factorization for face is used for editing the faces and magnifying the expression images. Learning based methods are used to synthesize expression

images in video to form a retrieved sequence, handle parallax without 3D reconstruction of that scene which is more effective for shaky motion removal.

Reviewing through tonal consistency techniques, the tonal alignment method fails to behave correctly when temporal uniformity is lost or damaged. Hence, color balancing technique is proposed to follow color changes from images. However the color correction is not able to reconstruct missing information in color balancing method. Therefore, color mapping come in view in which a precise global mapping cannot account for the entire image and a piecewise mapping is a feasible solution.

V. CONCLUSIONS

Expression transfer is achieved by blending a generic expression model with synthetically generated examples that better highlights person specific features. To set the identity we 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. Quantitative and qualitative evaluation shows that we significantly outperform previous approaches on achieving realistic, temporally coherent and accurate expression.

VI. ACKNOWLEDGMENT

I am indeed thankful to my guide **Prof S. S. Agrawal** for her able guidance and assistance to complete this paper; otherwise it would not have been accomplished. I extend my special thanks to Head of Department of Electronics & Telecommunication, **Dr. S. K. Shah** who extended the preparatory steps of this paper-work.

I am also thankful to the head & Principle of STES'S, SMT. Kashibai Navale College of Engineering, **Dr. A.V. Deshpande** for his valued support and faith on me.

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 and Gesture 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.

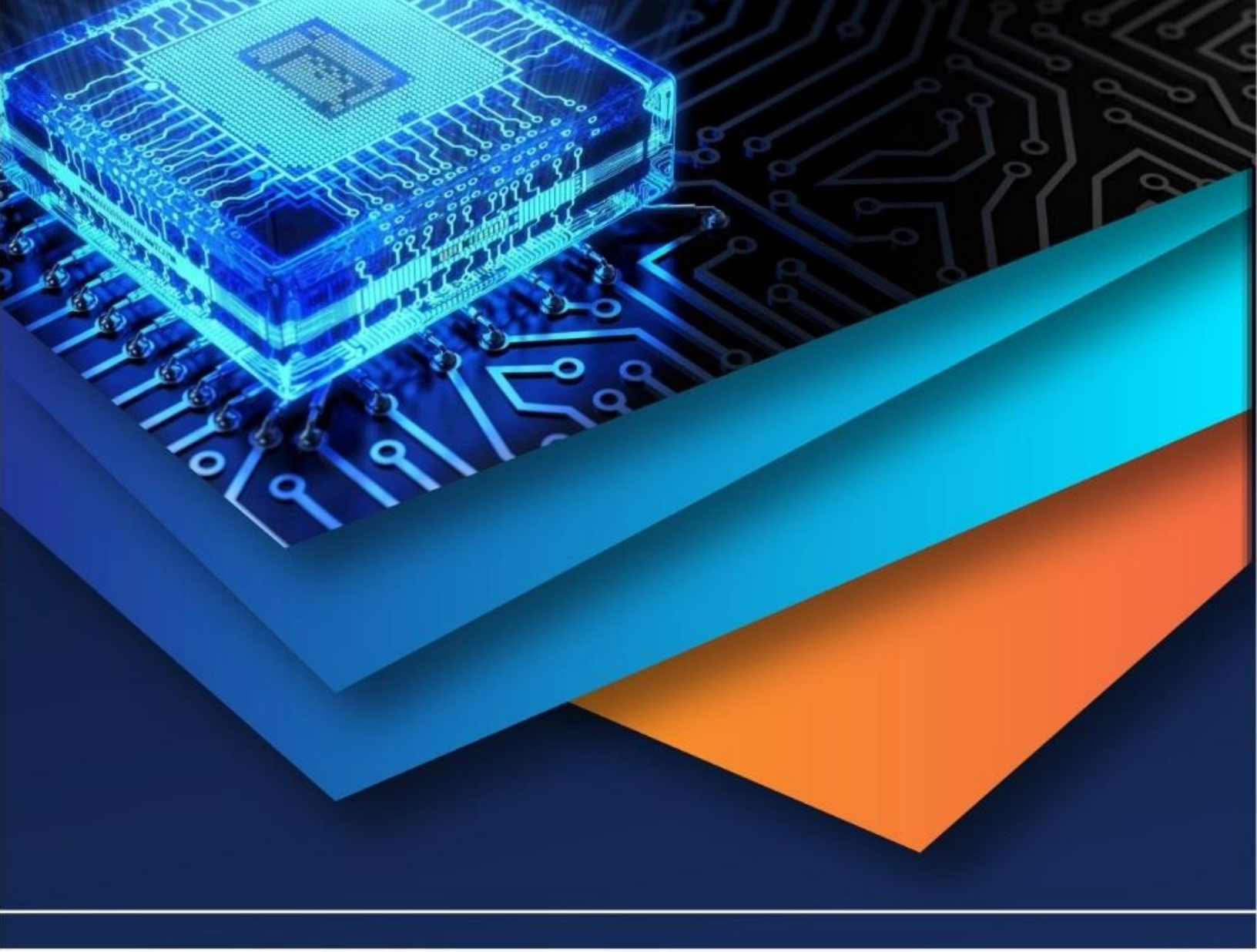
[7] F. Yang, L. Bourdev, E. Shechtman, J. Wang, and D. Metaxas, "Facial expression editing in video using a temporally-smooth factorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 861–868.

[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. Computer Vision and Pattern Recognition*, 2012, pp. 57–64.

[9] P. Litwinowicz and L. Williams, "Animating images with drawings," in *Proc. SIGGRAPH*, 1994, pp. 409–412.

[10] T. Beier and S. Neely, "Feature-based image metamorphosis," in *Proc. SIGGRAPH*, 1992, pp. 35–42.

[11] L. Williams, "Performance-driven facial animation," in *Proc. SIGGRAPH*, 1990, pp. 235–242.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY