3D MESH VIDEO RETRIEVAL: A SURVEY

Antonios Danelakis[a], *Theoharis Theoharis*[a,b], *Ioannis Pratikakis*[c]

[a]Department of Informatics & Telecommunications - National & Kapodistrian University of Athens, Greece
[b]Department of Computer & Information Science - Norwegian University of Science & Technology, Norway
[c]Department of Electrical & Computer Engineering - Democritus University of Thrace, Greece

ABSTRACT

This survey addresses methodologies for 3D mesh video retrieval including 3D mesh video action/motion retrieval and 3D mesh video facial expression recognition. They all involve retrieval procedures and, consequently, classification methods in order to identify similar actions/motions and facial expressions. The approaches are primarily categorized according to the 3D model representation that they use and their feature extraction and classification methods. Comparative data for the most promising methods is given, mainly on publicly available datasets.

***Index Terms* —** 3D mesh video, retrieval, survey.

1. INTRODUCTION

The recent availability of 3D mesh video and its potential applications are culminating research interest in the field. This paper is a survey of 3D mesh video retrieval algorithms, probably the first of its kind. It is focused on methodologies that use 3D meshes as frames representing the 3D video; throughout the paper, the term 3D video refers to 3D mesh sequences. We concentrate on 3D video action/motion retrieval and 3D video facial expression recognition. Biometric recognition algorithms are just a special case of retrieval and we focus on the shape descriptors used; the identification part of recognition corresponds to intra-class retrieval [1]. The presented methodologies are primarily categorized according to the 3D model representation that they use.

Table 1 illustrates the major publicly available 3D action and facial expression datasets. It should be noted that $BU-3DFE$ is a static 3D facial expression database, therefore, intermediate frames can be generated by interpolating among the frames representing four expression intensity levels.

DATASET	CONTEXT
BU-3DFE [2]	3D Facial Expressions
BU-4DFE [3]	3D Facial Expressions
Hi4D-ADSIP [4]	3D Facial Expressions
IXMAS [5]	3D Actions
i3DPost [6]	3D Actions / Facial Expressions

Table 1. Publicly available 3D action and facial expression datasets.

The typical operational pipelines comprise the stages shown in figures 1, 2.

2. 3D VIDEO ACTION/MOTION RETRIEVAL

An action is considered to be composed of motions. The first stage in the action/motion retrieval pipeline is a preprocessing step. This is optional and application dependent and is applied in order to simplify the task of subsequent stages. The feature extraction stage follows and finally retrieval, based on the features. Action/motion retrieval methodologies will be reviewed and categorized into volumetric-based, taking into account the whole volume of the 3D object, or surface-based, considering only the object's outer surface.

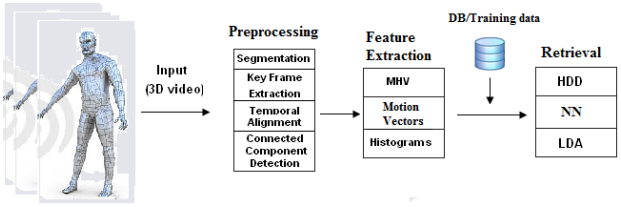

Figure 1. 3D video action/motion retrieval pipeline.

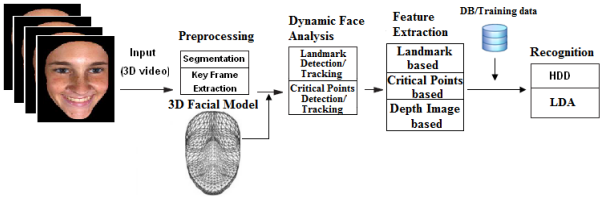

Figure 2. 3D video facial expression recognition pipeline.

2.1. Volumetric representation

Motion History Volume (MHV) is a spatio-temporal descriptor representing 3D video action/motion data. MHVs derived from different 3D sequences need to be aligned and compared. In [5], alignment and comparison is performed efficiently using Fourier transforms in cylindrical coordinates taken around the vertical axis. On the other hand, in [7], appropriate distance measures and parametric or non-parametric density estimators on specific manifolds are described. These measures are then used to learn class conditional densities for activity recognition, video-based face recognition and shape classification.

The analysis of volumetric data, obtained from the processing of 3D video sequences after enclosing the 3D object into one or more cylinders, is a very common action/motion retrieval methodology. In [8] a 3D cylindrical shape context is presented to capture the human body configuration for gesture analysis of volumetric data. Dynamics of motions are analyzed using Hidden Markov Models (HMM). In [9] a cylindrical wrapper of the human body is obtained in order to capture posture-dependent characteristics. Its output at each time instant constructs action feature matrices. For

action classification, the Dynamic Time Warp (DTW) approach is used. In [10] a set of cylinders of various sizes and orientations are created, enclosing various parts of the human body. Then these cylinders are used as 3D kernels in order to convolve over 3D volumes and find high response regions. Next, histograms holding the distribution of oriented cylinders are created. The histograms compose the feature vector and, for classification, DTW and HMM are used.

Circular layers intersecting the 3D object are also used for 3D video action/motion retrieval. In [11] the intersections of body segments with each layer are coded using enclosing circles. The circular features in all layers are used to generate a pose descriptor for each frame. The pose descriptors of all frames are combined to generate corresponding motion descriptors. For action recognition a simple nearest neighbor (NN) classifier is used. In [12], time series of circular FFT features, described in [5], are built. For classification, a Linear Discriminant Analysis (LDA) classifier is designed.

2.2. Surface representation

In [13] the 3D objects' vertices in each frame are used to generate distance histograms, based on D_2 distribution. The histograms of each frame are smoothed and motion segmentation is applied. Retrieval is performed using a Euclidean distance-based metric for comparing the motion segments of two 3D sequences. In [14] shape histograms use a spherical coordinate system and are constructed as follows. First, surface S of a model is isotropically scaled so that it lies within the unit sphere located at the origin and re-oriented per frame such that the direction of motion of its centroid is always along the Z axis. Then, S is represented by an implicit function and its shape histogram H is obtained. Kullback Leibler divergence ($K(h,e) = \sum_i h_i \log(h_i/e_i)$, where h and e are histograms of similar dimensions and i is an index over histogram bins), combined with a HMM, allows shape to be matched across noisy data.

Motion vectors are often used in order to capture temporal movement information, i.e. gestures. Motion vectors are based on 3D invariant statistical moments. In [15] the recognition of human gestures is performed by applying data fusion followed by motion vector extraction. A simple ellipsoid body model is fitted to incoming 3D data in order to capture which body part the gesture occurs in. Finally, a Bayesian classifier is employed to perform recognition over a small set of actions. In [16] a Harmonic Motion Context (HMC) is used to represent the extracted 3D motion vector fields efficiently. HMC is a view-invariant motion oriented 3D version of the shape context, which represents estimated 3D optical flow by embedded histograms of 3D Optical Flow (3D-HOF) in a spherical histogram. The resulting HMC descriptors are classified using normalized correlation. Motion analysis is also exploited in [17] and the concept of motion templates (MTs) is introduced. Using MTs, the essence of an entire class of logically related motions can be captured in an explicit and semantically interpretable matrix representation. In [18] a motion index tree, based on a hierarchical motion description, is constructed for a 3D library. The motion index tree serves as a classifier to determine the sub-library that contains the promising similar motions to a query. A nearest neighbor rule-based dynamic clustering algorithm is adopted to partition the library and construct the tree. The similarity between the query and a motion in the sub-library is calculated through elastic matching.

In [19] a distance function computing the similarity of the interaction of two characters using the spatial relationship of their body parts, is illustrated. For each interaction, a time varying graph structure based on the proximity of different joints is produced; the similarity of interactions is estimated by comparing the topology and Laplacian coordinates of different time-varying graphs.

Finally, in [20] both volumetric and surface representation is used. For the volumetric representation, Spin Images (SI) and Spherical Harmonics (SH) are employed to construct a frame-by-frame similarity matrix S_M. For the surface representation, Shape Distribution (SD) and Shape Histograms (SH) are used to construct S_M. In addition, two novel shape descriptors are proposed in order to construct the, so-called, multi-frame shape-flow similarity matrix and single-frame shape-flow matching similarity matrix. Temporal shape similarity is obtained by convolving the static shape similarity, reflected by the constructed similarity matrices, with a time filter. For the surface representation, In table 2, a summarization of 3D video action/motion retrieval methodologies is illustrated.

3. 3D VIDEO FACIAL EXPRESSION RECOGNITION

Although the current section deals with recognition, we concentrate on the descriptors and the 3D video representation used, which are related to the retrieval process. The pipeline for facial expression recognition begins with an optional preprocessing stage. Dynamic face analysis, which enables robust detection of facial changes, follows. Next, feature extraction, according to the detected facial changes, take place. Finally, recognition, based on the features, is performed; as mentioned earlier, recognition can be seen as a special case of retrieval. 3D video facial expression recognition methodologies will be reviewed and categorized based on the dynamic face analysis approach that they use. These approaches are temporal tracking of facial landmarks and critical points, mapping 3D facial scans onto a generic 3D face model or analyzing different facial surfaces in order to detect temporal facial changes.

3.1. Landmark tracking

In [21] a 2D tracker was employed and the facial model's projection was warped by 22 tracked feature points. The depth of a vertex was recovered by minimizing the distance between the model and the range data. Lipschitz embedding embeds the normalized deformation of the model in a low dimensional generalized manifold. For classification, a probabilistic expression model was learned on the generalized manifold. In [22] the composition of the descriptor and the classifier are the same as in [21] but in [22] the 2D face texture is generated using a conformal mapping and model adaptation algorithm. The proposed coarse-to-fine model adaptation approach between the planar representations was used and the correspondences are extrapolated back to the 3D meshes. In [24] 83 key vertices are tracked through the 3D video. Radial basis functions are used to adapt the generic model to the range facial model. Each adapted vertex is assigned one of eight possible primitive surface labels, by exploiting its principal curvature. Thus, a range model is represented by a label map composed of all vertices' labels in the facial region. LDA is used to project the range model to an optimal feature space. For classification, a HMM is used. In [23] an Active Shape Model (ASM)

METHOD	DATASET	NUMBER OF ACTIONS/MOTIONS	REPRESENTATION	FEATURES	CLASSIFIER	PERFORMANCE
Kohsia et al. [8]	Proprietary	7 motions	Volumetric	Cylindrical volume histograms	HMM	95.00% (Retrieval accuracy)
Weinland et al. [5]	IXMAS	11 actions	Volumetric	MHVs, FFT	LDA	93.33% (Retrieval accuracy)
Yamasaki et al. [13]	Proprietary	5 actions	Surface	Shape distributions	NN	83.78% (Retrieval accuracy)
Turaga et al. [7]	IXMAS	11 actions	Volumetric	MHV	NN	98.78% (Retrieval accuracy)
Veeraraghavan et al. [12]	IXMAS	11 actions	Volumetric	Circular FFT	NN (DTW)	98.18% (Retrieval accuracy)
Pehlivan et al. [10]	IXMAS	5 actions	Volumetric	Distribution histograms	HMM, NN (DTW), SVM	83.89% (NN), 80.74% (HMM) (Retrieval accuracy)
Holte et al. [16]	i3DPlot	8 actions	Surface	3D motion vectors	Normalized correlation	92.19% (Retrieval accuracy)
Pehlivan et al. [11]	IXMAS	13 actions	Volumetric	Circular body layer projections	NN	88.63% (Retrieval accuracy)
Canton et al. [15]	Proprietary	8 actions	Surface	Invariant statistical moments	Bayes	89.30% (Retrieval accuracy)
Feng et al. [18]	Proprietary	12 actions	Surface	Motion index tree	NN	$\approx$ 98.00% (from P-R diagram)
Müller et al. [17]	Proprietary	11 motions	Surface	Motion templates	NN (DTW)	$\approx$ 84.80% (from P-R diagram)
Tang et al. [19]	Proprietary	285 motions	Surface	Time-varying spatial graphs	NN (DTW)	$\approx$ 96.00% (from P-R diagram)
Huang et al. [20]	Proprietary	28 motions	Volumetric/Surface	Shape histograms	NN	$\approx$ 96.00% (from ROC diagram)
Kilner et al. [14]	Proprietary	6 motions	Surface	Shape histograms	HMM	79.49% (601/756 True positive hits)
Pierobon et al. [9]	Proprietary	3 actions	Volumetric	Cylindrical volume histograms	NN (DTW)	N/A

Table 2. Overview of research work on 3D video action/motion retrieval.

METHOD	DATASET	NUMBER OF EXPRESSIONS	AUTOMATIC	3D FACE ANALYSIS	CLASSIFIER	CLASSIFICATION ACCURACY
Chang et al. [21]	Proprietary	6	NO	Landmark tracking	Bayes	N/A
Rosato et al. [22]	BU-4DFE	4	YES	Landmark tracking	Bayes	85.90%
Tsalakanidou et al. [23]	Proprietary	4	YES	Landmark tracking	FACS	85.00%
Sun et al. [24]	BU-4DFE	6	NO	Landmark tracking	HMM	90.44%
Canavan et al. [25]	BU-4DFE	6	YES	Landmark tracking	SVM	86.30%
Berretti et al. [26]	BU-4DFE	3	YES	Critical point tracking	HMM	83.75%
Yin et al. [27]	BU-3DFE	6	NO	3D facial model-based	LDA	80.20%
Sandbach et al. [28]	BU-4DFE	3	YES	3D facial model-based	HMM	81.93%
Sandbach et al. [29]	BU-4DFE	6	YES	3D facial model-based	GB & HMM	64.59%
Tianhong et al. [30]	BU-4DFE	6	YES	3D facial model-based	NN	74.63%
Vuong et al. [31]	BU-4DFE	3	YES	Facial surface-based	HMM	92.22%
Hassen et al. [32]	BU-4DFE	6	YES	Facial surface-based	LDA & MRF	93.21%

Table 3. Overview of research work on 3D video facial expression recognition.

is built in order for 81 3D facial landmarks to be selected. The ASM is then fitted onto the data using the gradient information in the neighborhood of each landmark. The feature vectors combine geometric information of the landmarks and the statistics on the density of edges and curvature around the landmarks according to the Facial Action Coding System (FACS), where facial changes are described in terms of 44 Action Units (AU). Finally, in [25], 3D landmark tracking is applied and the tracked landmarks are used for curvature-based feature extraction. For classification, a SVM classifier is exploited.

3.2. Critical point tracking

In [26] automatic selection of points on the nose, eyes and mouth using z-buffers takes place. A face in a 3D frame is represented by computing and averaging distances between the detected facial points. These distances are then normalized, quantized and summed in a final descriptor. HMM is used for system training and classification. In table 3, a full summarization of 3D video facial expression recognition methodologies is given. For each method, the top classification accuracy reported is shown.

3.3. Facial surface-based

In [31] facial level curves on the Z axis are created, at different heights h. Every facial point at height h belongs to the corresponding curve. Comparison between same level curves leads to a distance vector (descriptor) for each frame. The descriptors corresponding to individual frames are combined to create an augmented vector. PCA and LDA are used to decrease the dimensionality of the descriptor and a HMM is employed for classification.

3.4. 3D facial model-based

In [27] a tracking 3D model for estimating motion trajectories, which are used to construct a spatiotemporal descriptor called facial expression label map (FELM), is proposed. The tracking model is first aligned to the 3D face scan, and then deformed to fit the target scan by minimizing an energy function. The FELM vector and the motion vector are concatenated to form the descriptor, which becomes the input to a LDA classifier. In [28] free form deformations are used in order to find a vector field reflecting facial motion. Next, 2D feature extraction takes place for every frame. All derived features are concatenated into one feature vector per frame in the image sequences, and these are used for classification. For classification, a HMM is used. In [29] a similar approach is adopted. This approach focuses on the facial regions which present the greatest amount of motion. The classification process in enriched by using GentleBoost (GB) classifiers in addition to HMM. In [30] a mesh matching procedure, based on facial vertex correspondence, is applied. Procrustes analysis is used to determine the correspondence transformation. To construct the final descriptor, the pixels of an image are labeled by thresholding each pixel's neighborhood with the center value. The results are translated into binary numbers, which codify local patterns of different types and are accumulated in a histogram over a predefined region. Temporal evolution is also considered. This histogram essentially becomes the descriptor of the region and the whole image can be described by a concatenation of such histograms. In [32] a new deformation vector field descriptor is proposed. The facial surfaces are represented by a set of parameterized radial curves emanating from the tip of the nose, which defines the novel descriptor. Then, a LDA-based transformation is used for dimensionality reduction. Finally, the Multiclass Random Forest (MRF) learning algorithm is exploited for the classification process.

4. CONCLUSIONS / DISCUSSION

State-of-the-art methodologies aimed at 3D video retrieval were reviewed. 3D video is represented as a sequence of 3D meshes.

The model representation in action/motion retrieval can be either volumetric or surface-based. It is remarkable that methodolo-

gies using MHV or other frequency domain descriptors seem to perform better. In addition, all action/motion retrieval methodologies are automatic. We point out that landmark tracking-based methodologies, which could be very promising, have not been found in the literature.

On the other hand, the model representation for facial expression recognition is always surface-based. Dynamic face analysis is an additional step in the recognition pipeline which exploits the fact that human facial motions are discrete and well studied and facilitates feature extraction. Mapping 3D facial scans onto a generic face model is very often used for this purpose. At the recognition stage, similar techniques to action/motion retrieval are used, with HMM having a prominent role. We note that there are some methodologies in this area which are not fully automated.

In the near future, it is expected that more actions/motions and facial expressions, including arbitrary expressions, are going to be taken into account. The newly developed database in [4] points in this direction.

Acknowledgement

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS.

5. REFERENCES

[1] Passalis, G. and Kakadiaris, I. A. and Theoharis, T., "Intraclass retrieval of nonrigid 3D objects: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 218–229, 2007.

[2] Y. Lijun, W. Xiaozhou, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *IEEE Proc. FGR '06*, 2006, pp. 211–216.

[3] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *IEEE Proc. FG '08*, 2008, pp. 1–6.

[4] B. Matuszewski, W. Quan, L.K. Shark, A.S. McLoughlin, C.E. Lightbody, H.C.A. Emsley, and C.L. Watkins, "Hi4D-ADSIP 3D dynamic facial articulation database," *Elsevier Image and Vision Computing*, pp. 1–15, 2012 (doi: 10.1016/j.imavis.2012.02.002).

[5] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Elsevier Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[6] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3D human action/interaction database," in *Proc. CVMP '07*, 2009, pp. 159–168.

[7] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision.," in *IEEE Proc. CVPR '09*, 2008.

[8] K. S. Huang and M. M. Trivedi, "3D shape context based gesture analysis integrated with tracking using omni video array," in *IEEE Proc. CVPR '05*, 2005, pp. 80–88.

[9] M. Pierobon, M. Marcon, A. Sarti, and S. Tubaro, "3D body posture tracking for human action template matching," in *IEEE Proc. CVPR '05*, 2006, vol. 2, pp. II501–II504.

[10] S. Pehlivan and P. Duygulu, "3D human pose search using oriented cylinders," 2009, IEEE Conf. ICCV '09, pp. 16–22.

[11] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *Elsevier Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 140–151, Feb. 2011.

[12] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa, "Rate-invariant recognition of humans and their activities," *IEEE Trans. on Image Processing*, vol. 18, no. 6, pp. 1326–1339, June 2009.

[13] T. Yamasaki and K. Aizawa, "Motion segmentation and retrieval for 3D video based on modified shape distribution," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 211–222, 2007.

[14] J. Kilner, J. Y. Guillemaut, and A. Hilton, "3D action matching with key-pose detection," in *Search in 3D and Video (S3DV)*, 2009, pp. 1–8.

[15] C. Canton-Ferrer, J. R. Casas, and M. Pardas, *Human model and motion based 3D action recognition in multiple view scenarios*, pp. 14–18, 14th European Signal Processing Conference EUSIPCO. 2006.

[16] M. B. Holte, T. B. Moeslund, N. Nikolaidis, and I. Pitas, "3D human action recognition for multi-view camera systems," 2011, IEEE Proc. 3DIMPVT '11, pp. 342–349.

[17] M. Müller and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data," 2006, ACM SIGGRAPH/Eurographics Proc. SCA '06, pp. 137–146.

[18] F. Liu, Y. Zhuang, F. Wu, and P. Yunhe, "3D motion retrieval with motion index tree," *Elsevier Compuer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 265–284, 2003.

[19] J. Tang, J. Chan, H. Leung, and T. Komura, "Interaction retrieval by spacetime proximity graphs," in *Computer Graphics Forum*, 2012, vol. 31, pp. 745–754.

[20] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3D video sequences of people," *Springer International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362–381, 2010.

[21] Y. Chang, M. B. Vieira, M. Turk, and L. Velho, "Automatic 3D facial expression analysis in videos," in *IEEE Workshop AMFG '05*, 2005, pp. 293–307.

[22] M. Rosato, X. Chen, and L. Yin, "Automatic registration of vertex correspondences for 3D facial expression analysis," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2008, pp. 1–7.

[23] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Elsevier Pattern Recognition*, vol. 43, no. 5, pp. 1763–1775, May 2010.

[24] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Springer Proc. ECCV '08: Part II*, 2008, Springer Proc. ECCV '08, pp. 58–71.

[25] C. Shaun, S. Yi, Z. Xing, and Y. Lijun, "A dynamic curvature based approach for facial activity analysis in 3D space," in *SISM '12*, 2012, pp. 14–19.

[26] S. Berretti, A. Del Bimbo, and P. Pala, "Real-time expression recognition from dynamic sequences of 3D facial scans," in *EU Workshop on 3D Object Retrieval*, 2012, pp. 85–92.

[27] L. Yin, X. Wei, P. Longo, and A. Bhuvanesh, "Analyzing facial expressions using intensity-variant 3D data for human computer interaction," in *in Proc. ICPR '06*, 2006, pp. 1248–1251.

[28] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *IEEE FG '11*, 2011, pp. 406–413.

[29] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Elsevier Image and Vision Computing*, 2012 (doi: 10.1016/j.imavis.2012.01.006).

[30] F. Tianhong, X. Zhao, K. S. Shishir, and I. A. Kakadiaris, "4D facial expression recognition," in *ICCV '11*, 2011, pp. 1594–1601.

[31] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *IEEE FG '11*, 2011, pp. 414–421.

[32] D. Hassen, B. Boulbada, D. Mohamed, S. Anuj, and B. Stefano, "3D dynamic expression recognition based on a novel deformation vector field and random forest," in *ICPR '12*, 2012.