iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** IX **Month of publication:** September 2014

DOI:

www.ijraset.com

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

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Image-Based 3D Reconstruction

Tushar Mehmi[1], Vishal Yadav[2], Upender Yadav[3]

Dronacharya College of Engineering, India

Abstract: ***We present various method for the reconstruction of Image based structures. In this paper we combine methods from the field of computer vision with surface editing techniques, method to reconstruct 3D character models from video, design of an interactive image-based modeling tool that enables a user to quickly generate detailed3D models with texture from a set of calibrated input images, Estimating photo-consistency is one of the most important ingredients for any 3D stereo reconstruction technique that is based on a volumetric scene representation, reconstructing watertight triangle meshes from arbitrary unoriented point clouds.***

Keyword: Reconstruction, tracking, Surfel fitting etc.

I. INTRODUCTION

A. Character Reconstruction and Animation from Uncalibrated Video

Recent techniques for image-based 3D character reconstruction have been able to create and animate virtual character models of very high quality However, most approaches require accurately calibrated systems with multiple synchronized cameras, exact silhouette information, or additional hardware like range sensors. We envision 3D character reconstruction as a simple image processing tool, which allows to reconstruct a virtual model of reasonable quality by simply filming different views of a person with a hand-held camera.

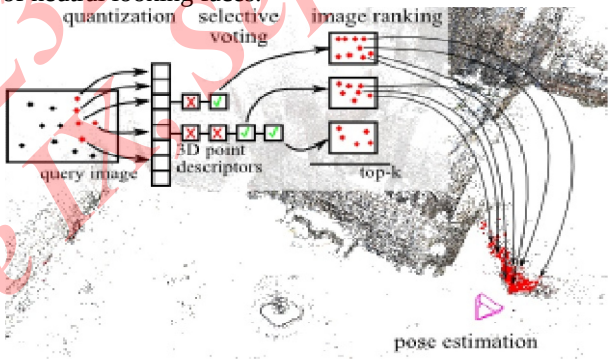
In this paper we show that character reconstruction from a single, un calibrated video showing a person in articulated motion is nevertheless possible. Our main conceptual contribution is the transformation of this problem into a synchronized multi-view setting by pose synchronization of the character.

B. Marker less Reconstruction of Dynamic Facial Expressions

The acquired motion data can be used to create animations for movies, computer games, or humanoid avatars which can be utilized in scientific as well as industrial applications.

In this paper, we exploit methods from computer vision and mesh editing to compute a dense motion field for facial animations from synchronized video streams. The motion field is represented by a predefined face template whose vertices move in time according to the underlying scene flow.

Since we use a predefined face template which is fitted to an individual face, we immediately obtain the correspondences between all acquired reconstructions. Our predefined face template is a simple morphable model whose low-dimensional set of parameters is able to control the shape of neutral looking faces.



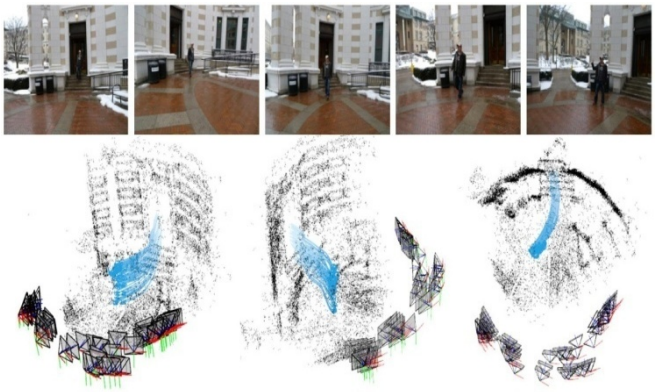


Fig.: Showing examples of 3D image reconstruction

II. OVERVIEW

- Character Reconstruction and Animation from Uncalibrated Video

Research on image-based 3D character reconstruction and animation has primarily focused on controlled acquisition setups with multiple synchronized video streams. One class of techniques reconstructs a character model by computing the visual hull from several silhouettes which can then be refined for a more faithful reconstruction or used, e.g. for tracking and pose estimation. These methods require an accurate calibration and a segmentation of the character from the background. Alternatively, a pre-defined human body model can be fitted to the silhouette of a human character. Based on the SCAPE model or other articulated character templates it is even possible to model and animate fine scale shape changes .However, these methods cannot easily be adapted to shapes which differ considerably from the underlying model.

Some techniques reconstruct non-rigid surfaces from monocular input video. However, their specific constraints and the lack of occlusion handling render these methods inapplicable for a practical system for articulated character reconstruction. An alternative is to first estimate the character pose and use this information to facilitate reconstruction. But in order to achieve the accuracy required for 3D reconstruction, further information such as a segmentation is still necessary. Existing methods for camera calibration such as Sf M or model-based techniques are not suitable in our setting with a non-static camera and an independently moving character. Since one of our primary aims is to review the method applicable to a variety of character types and input videos, the use of semi-automatic camera and pose estimation is done.

The goal of this method is to reconstruct a 3D model from a single input video of a person. This problem setting is highly ill-posed: Due to articulated motion each frame of the video might show the person in a slightly different pose. Moreover, we do not assume any prior video processing such as camera calibration or image segmentation. We show that, by utilizing a generic 3D character shape template and a database of 3D motion data, a single video of a person can be effectively converted into a temporally synchronized multi-view setup, a process which we refer to as pose synchronization. The converted video then allows for an image-based 3D reconstruction of the character shape.

The first step is to fit a generic shape template M to some input frame I_t of the video. This shape template is a 3D triangle mesh with an embedded skeleton.

Given a skeleton pose τ ,e.g. from a motion database, the template can be deformed accordingly $M \rightarrow M^t$ using skeleton-driven animation techniques. In order to fit his template model to the character in image I_t we first estimate the approximate character pose $\tau(t)$ and a projection Pt from the 3D motion data to the 2D image space based on user-selected joint positions. The deformed template $M^{\tau(t)}$ is then projected into It (and its silhouette vertices are aligned with the character's silhouette. The resulting reference shape St, which consists of all projected front-facing triangles, provides an initial mapping from the generic shape template M in 3D to the character in the 2D image I_t.

For a 3D reconstruction, dense 2D correspondences between different views of the character are required. This is a challenging problem due to frequent occlusions and the dynamic nature of our input videos.

We compute these correspondences by a novel mesh-based approach which is able to track the reference shape St through a subsequence of the video $\{I_s,\ldots,I_t,\ldots,I_r\}$. Occlusions are resolved by utilizing the depth information of the different layers of front-facing triangles in St.

The result is a shape sequence $St = \{S_s,\ldots,S_t,\ldots,S_r\}$ where corresponding vertices in St have consistent positions on the character. The main idea of pose synchronization is to convert the video sequence into a synchronized multi view setup by eliminating the pose differences. We achieve this by computing an as-rigid-as-possible deformed shape S_j^τ for each tracked shape $S_j \in \mathcal{S}_t$, according to a common pose τ from the motion database. By combining several partial updates from different sub sequences of the input video, a consistent

ISSN: 2321-9653

3D model of the filmed character can be created and animated with the available motion capture data.

- Shape Tracking

In order to track the character shape we experimented with a variety of standard approaches such as feature tracking, correspondence estimation, and optical flow. These types of approaches revealed a number of fundamental restrictions. For instance, tracking rectangular windows centred, e.g., at the mesh vertices of St, or optical flow has the drawback that it is difficult to handle occlusions and discontinuities at silhouettes of the limbs and body. Techniques for correspondence estimation generally do not provide sufficiently dense matches. These issues become even more severe for the limited character size at standard video resolutions. Moreover, we have to keep track of the complete limbs of a character even if they are partially occluded. We therefore developed a novel mesh-based tracking approach, which exploits the depth information in a layered reference shape St in order to resolve occlusions and to keep track of occluded limbs.

Given two successive images I_j and $I_{j+1}$and a shape S_j in image I_j, our goal is to compute a displacement field attached to the vertices of S_j, i.e. a displacement di for each vertex vi of S_j. The vertices $\mathbf{v}_i'$ of S_{j+1} then become $\mathbf{v}_i' := \mathbf{v}_i + \mathbf{d}_i$. Each triangle face f_k of S_j, together with the respective transformed face f_k' of S_{j+1}, defines an affine transformation A_k: $R^2 \rightarrow R^2$between the images I_j and I_{j+1}. We formulate the matching process as a global optimization problem that minimizes the sum of triangle errors. The per-triangle error for each pair (fk, f'k) of corresponding triangles is computed by summing over the squared intensity differences of the respective image area sin I_j and I_{j+1}. The desired displacement field is then the minimum of the objective function

$$E_{\text{data}} = \sum_{f_k \in S_j} \sum_{\mathbf{p} \in \Omega_k} \left(I_j(\mathbf{p}) - I_{j+1}(A_k(\mathbf{p})) \right)^2$$

Where $\Omega_k \subset \mathbf{N}^2$ denotes the set of image pixels covered by triangle f_k in image I_j. Coherent tracking and robustness to image noise is ensured by enforcing an additional term

$$E_{\text{smooth}} = \sum_{i \in V(S_j)} \frac{1}{\omega_i} \sum_{j \in N_i} \omega_{i,j} \|\mathbf{d}_i - \mathbf{d}_j\|^2,$$

which imposes rigidity on the tracked mesh. V (S_j) denotes the set of vertices of the shape S_j and N_i denotes the 1-ring neighbours of vertex i. We chose the standard chordal weights $\omega_{i,j} := \|\mathbf{v}_i - \mathbf{v}_j\|^{-1}$, $\omega_i := \sum_{j \in N_i} \omega_{i,j}$.

The complete objective function then is

$$E = E_{\text{data}} + \lambda E_{\text{smooth}},$$

which is minimized using the Levenberg-Marquardt algorithm to determine the vertex displacement field between pairs of successive images.

- Pose Synchronization

Articulated character motion within a tracked image sequence leads to global shape distortions, However, assuming continuous character motion without too large shape changes, a single video sequence can be converted such that it approximates a temporally synchronized multi-view setup by synchronizing all
the tracked shapes. First, camera projections P_j are computed for the shapes $S_j \in \mathcal{S}_t$. Since the 2D skeleton joints are pulled along with the shapes during the mesh tracking, one generally has to do only minor adjustments to place the joint positions at their approximate locations. The best matching common pose $\top$ for all shapes

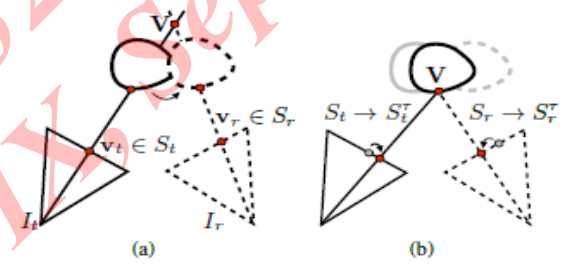


Figure (a) Triangulation of 2D correspondences $\mathbf{v}$ between views of a non-rigid scene results in wrong reconstructions $\mathbf{V}$. (b) Our pose synchronization transforms the correspondences into a rigid setting for a corrected estimate.

In practice we do the synchronization only for the reference shape St and the two shapes $S_t$and S_r at the boundaries of the tracked video interval. In general the corresponding views have the largest baseline and hence result in the most stable 3D reconstruction. The effective solution to this problem is the computation of interpolated 2D vertex positions for each shape similar to the concept of epi pole consistent camera weighting. Suppose we have two shapes S_i and S_j from two different shape sequences in the same image. Then some vertices of the template model M are likely to be visible in both shapes S_i and S_j. Let $v_i$and v_j be two corresponding 2D positions of a particular vertex. Then the updated vertex position $v^*$is computed by a weighted contribution

$\mathbf{v}^\star = \sum_k \omega_k \mathbf{v}_k$. Otherwise we weight the contribution to the position using $\tilde{\omega}_k = (|k-t|+1)^{-\beta}$.

The final weight is normalized by

$$\omega_k = \tilde{\omega}_k / \sum_l \tilde{\omega}_l.$$

However, by initializing the vertex positions in a new shape sequence with tracked vertex positions from a previously generated sequence, it is possible to compute partial reconstructions with pixel accurate surface points even for such complex models.

- Markerless Reconstruction of Dynamic Facial Expressions

Each of the cameras record images at 30 FPS with a resolution of 640 × 480 (Figure above shows the rig, together with four of these images in the middle of a sequence). If necessary, the frame rate could be increased to 60 FPS by triggering the cameras externally.

The first image of each sequence shows the face in its neutral pose. In order to be able to track facial movement we need to reconstruct the face seen in the first frame. Independently optimizing point depth values using *Surfel Fitting* would produce a point cloud of arbitrary size. This drastically increases to robustness of the Surfel Fitting because of the good initial Surfel parameters. One requirement of our system is that inter-subject correspondences have to be maintained. This becomes possible by using a face template, containing a fixed number of vertices, which is in a one tone correspondence with the vertices of the morphable model.

To achieve correspondences between the vertices of the template and the captured face which are maintained while initial face template is deformed during the whole sequence. Combination of mesh modeling techniques with multi view stereo reconstruction: 2D image-samples, placed in the first image of every view, are tracked over the entire sequence .In order to establish temporal correspondences between successive frames, we use a 2D mesh based tracking approach. Simple feature tracker like the KLT tracker often have the problem that features slide past each other, since their displacements ore optimized independently of their local neighbourhood. In the proposed 2D mesh tracking we can control for the global smoothness of the produced mapping, to prevent fold overs. Shooting rays through an image-sample of the first image hits the initial face template and thereby defines an anchor point in 3D. At each step, the tracked image-samples are reconstructed using the Surfel Fitting approach. Together with its anchor point lying on the surface of the face template, a successfully reconstructed image-sample will provide a constraint which is used in the modeling step to deform the face template. The reviewed modelling step has two advantages: First, if the Surfel Fitting does not succeed the face template can still be deformed using surrounding successfully reconstructed Surfels . Second, since the tracked face template provides good initial solutions, Surfel Fitting becomes much more robust.

- Surfel Fitting

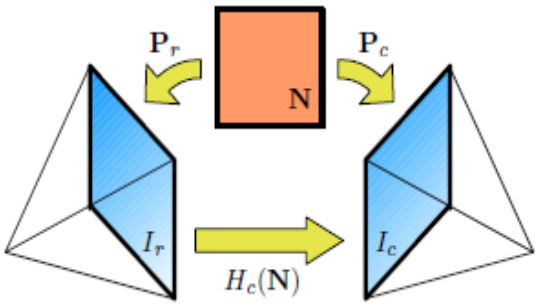


Figure The 3D plane together with the image projection matrices define a homography H which maps image points from I_r to points in the image I_c.

The Surfel's associated plane defines a homography which maps pixels from a calibrated reference image I_r to a calibrated comparison image I_c. Surfel Fitting optimizes the parameters of the plane by minimizing pixel intensity difference between the reference and comparison images.

ISSN: 2321-9653

Given the input plane defined by the initial position $\mathbf{p} \in \mathbb{R}^3$ and normal $\mathbf{n} \in \mathbb{R}^3$, a reference image, and a set of comparison images for the plane are defined by:

$$\begin{aligned} \mathbf{ref}(\mathbf{p}) &= I_r & r \in \{1,\dots,C\} \\ \mathbf{comp}(\mathbf{p}) &= \{I_{c_1},\dots,I_{c_k}\} & c_1,\dots c_k \in \{1,\dots,C\} \end{aligned}$$

Where C is the number of views. The reference image can for example be chosen as the image where viewing direction and the vertex normal are closest to parallel. In all the presented steps of the tracking workflow, we have a good initial closed surface. Thus, the set of comparison images can be obtained by a simple visibility test using the OpenGL z-Buffer. Since Surfel Fitting is used to reconstruct the initial face, as well as to track facial movements we will describe this approach in the next section. The advantage of this algorithm is its simplicity since each Surfel can be optimized independent of its local neighbourhood.

Let the projection matrices for the reference image and the comparison image be :

$$\mathbf{P}_r = [\mathbf{Q}_r|\mathbf{q}_r] \text{ and } \mathbf{P}_c = [\mathbf{Q}_c|\mathbf{q}_c]$$

Without loss of generality, we can transform the scene by a matrix $\mathbf{B}$ such that $\mathbf{P}'_r = \mathbf{P}_r\mathbf{B} = [\mathbf{Id}_3|0]$. Together with its normal $\mathbf{n}$ we define a plane at point $\mathbf{p}$ as $\mathbf{N}^T = [\mathbf{n}^T, \delta]$, with $\delta = -\mathbf{p}\cdot\mathbf{n}$. This determines a homography

$$\begin{aligned} H_c(\mathbf{N}) &= (\delta\mathbf{Q}_c - \mathbf{q}_c\mathbf{n}^T)(\delta\mathbf{Q}_r - \mathbf{q}_r\mathbf{n}^T)^{-1} \\ &= (\delta\mathbf{Q}'_c - \mathbf{q}'_c\mathbf{n}^T) \end{aligned}$$

which maps pixels $\hat{\mathbf{p}} \in \mathbb{R}^2$ from the reference image to the comparison image (Figure 2). The objective is to find new plane parameters which minimize the energy function

$$E_c(\mathbf{N}) = \sum_{\hat{\mathbf{p}}\in\Omega} \left(I_r(\hat{\mathbf{p}}) - I_c(H_c(\mathbf{N})\hat{\mathbf{p}})\right)^2$$

Fig. describing the formulaes for surfel fitting.

Where Ω is a square region in the reference image around the projected vertex $P_r p$. Notice that the final energy takes all comparison images into account and can be expressed as

$$E(\mathbf{N}) = \sum_c E_c.$$

After minimizing the energy function, the 3D position can be obtained by shooting a ray through the centre of Ω and computing the intersection with the optimized plane. Occasionally, due to noise in the images or badly textured parts in human faces, this process does not succeed at every vertex.

- Mesh Tracking in 3D

The objective of the algorithm described in this section is to find a deformation of the face template for every frame, such that the highly detailed movements of the captured face are tracked by the template face. To achieve this, we generate image samples in every view and track them in time. Using the Surfel Fitting , these image-samples are reconstructed in 3D for every frame. Finally, these reconstructed 3D points are used to deform the template mesh. Supers ampling the triangles of the meshes in the first frame of the sequence generates 2D points that have bary centric coordinates w.r.t. the triangle they are placed in.

For every view c this yields a set of points $\hat{\mathbf{p}}^1_{i,c}$ for the first frame, where a point can uniquely be identified by its index i and the view c it was put in. The mapping π, which is defined by the deformation of a 2D mesh from one frame to the following, allows us to displace the image samples and thereby track them through the whole sequence of a single view. This produces sequences of points $\hat{\mathbf{p}}^1_{i,c}$.

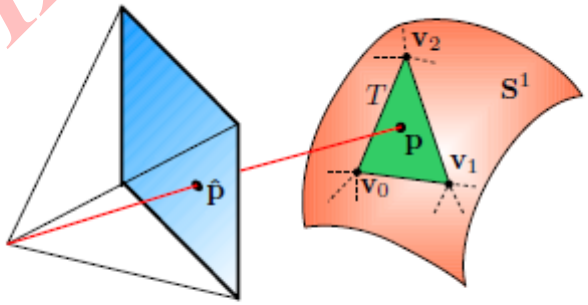


Figure An anchor point of a image-sample $\hat{\mathbf{p}}$ is obtained by the intersection of a ray through $\hat{\mathbf{p}}$ with the initial face template $\mathbf{S}^1$.

It is worth mentioning that this procedure can also help improve the estimated surface $S^1$of the first frame. At the end of the process described in, a new point cloud can be extracted by Surfel Fitting. For each Surfel, we can compute an anchor point as the closest point on S^1w.r.t. the Surfel. These pairs can then be used to deform the face template S^1, as described above. The deformed surface does not lie in the space spanned by the morphable model and is used as the input surface for all subsequent steps of the pipeline.

- Deformation of the template mesh.

In order to deform the mesh we treat as handles which drag the surface S^1. We define two objective functions. The first function measures the squared distance between the Laplace vectors of $S^1$and those of the deformed surface S^f

$$E_L = \sum_{\mathbf{v} \in \mathbf{S}} \| \Delta \mathbf{v}^f - \Delta \mathbf{v}^1 \|^2$$

Here, Δ denotes the discrete Laplace operator using the cotangent weights evaluated on the surface S1. The second function penalizes large deviation of the anchor point from the reconstructed point and can be denoted as

$$E_C = \sum_{\mathbf{p} \in \mathbf{Succ(p)}} \| \mathbf{p} - \mathbf{anchor(p)} \|^2$$

where the anchor point anchor $(\mathbf{p}_{i,c}^f)$ is calculated by interpolating the vertices of the triangle *T* associated with $\hat{\mathbf{p}}_{i,c}^1$.

using the pre computed bary centric coordinates:

$$\mathbf{anchor}(\mathbf{p}_{i,c}^f) = \sum_{\mathbf{v}_t \in T} \gamma_t \cdot \mathbf{v}_t^f$$

To obtain the new vertex positions of a mesh S^f, we reviewed $E = E_L + \lambda E_C$ in the least-squares sense and repeat the whole procedure for the next frame f + 1.

III .CONCLUSION

In this paper we presented a way to overcome the problem of energy function which minimizes only a small local image region which is considered by using a simple morphable model of neutral faces to estimate the more global appearance of the face seen in the first image. This generates a surface similar to the one being reconstructed, which strongly increases the robustness of the Surfel Fitting. First, the temporal tracking does not take the neighbouring features into account. Because of that, it often produces trajectories which slide past each other than these fold overs inducing high distortions in the tracked face template. To correct this, we chose the proposed 2D mesh tracking because we can control the global smoothness and prevent the features from sliding. Second, tracking features between views produces only a very sparse set of 3D features, which do not provide enough constraints for the modeling step to get reliable results. In our proposed method, we distribute a large set of (redundant) image-samples, so we are able to omit wrongly reconstructed image-samples but still end up with a large set of constraints for the modeling phase. Since each image samples considered independently, the 3D reconstruction is simple. Combining it with a simple modeling approach which fulfils each constraint in the least squares sense.

We reviewed in this paper a novel, semi-automatic approach to reconstruct and animate 3D character models from uncalibrated, monocular video. The main technical contributions discussed here is algorithm for pose synchronization to compensate for articulated character motion. We demonstrated that it is possible to produce reasonable reconstruction and animation results for a range of different character types. However, we did not intend to compete with the very high quality possible with more complex and constrained capture systems. There is, however, still some room for further improvement: The currently used objective function based on the SSD of intensity values with photometric normalization works well for moderate lighting changes but is known to have problems with stronger changes like specular reflections. We believe that our system nicely complements existing work and provides a first basis for making 3D character reconstruction from general video a simple image processing task. This opens up entirely new opportunities and applications such as, for example, the reconstruction of historical characters from film archives or the creation of realistic 3D avatars for home users.

REFERENCES

[1] Seitz, S.M., Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. In: CVPR. (1997) 1067–1073

[2] Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International Journal of Computer Vision 38 (2000) 199–218 190 A. Hornung and L. Kobbelt

[3] Esteban, C.H.: Stereo and Silhouette Fusion for 3D Object Modeling from Uncalibrated Images Under Circular Motion. PhD thesis, Ecole Nationale Sup´erieure des T´el´ecommunications (2004)

[4] Vogiatzis, G., Torr, P., Cipolla, R.: Multi-view stereo via volumetric graph-cuts. In: CVPR. (2005) 391–398

[5] Sinha, S., Pollefeys, M.: Multi-view reconstruction using photo-consistency and exact silhouette

constraints: A maximum-flow formulation. In: ICCV. (2005)

[6] Slabaugh, G.G., Schafer, R.W., Hans, M.C.: Image-based photo hulls for fast and photo-realistic new view synthesis. Real-Time Imaging 9 (2003) 347–360

[7] Li, M., Magnor, M., Seidel, H.P.: Hardware-accelerated rendering of photo hulls. Computer Graphics Forum 23 (2004) 635–642

[8] Bonet, J.S.D., Viola, P.A.: Roxels: Responsibility weighted 3D volume reconstruction. In: ICCV. (1999) 418–425

[9] Z´yka, V., S´ara, R.: Polynocular image set consistency for local model verification. In: Workshop of the Austrian Association for Pattern Recognition. (2000) 81–88

[10] Broadhurst, A., Drummond, T., Cipolla, R.: A probabilistic framework for space carving. In: ICCV. (2001) 388–393

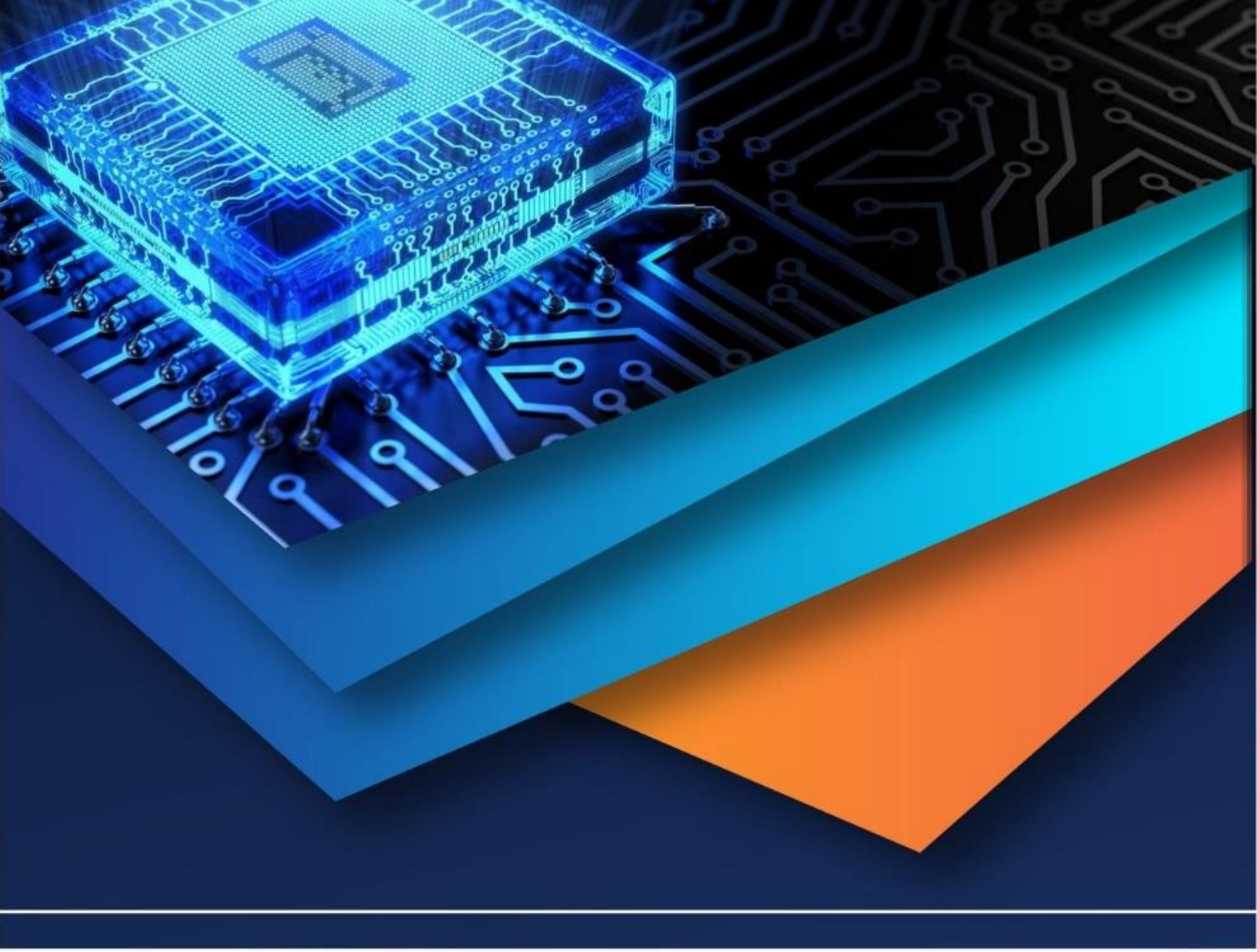
[11] Stevens, M.R., Culbertson, W.B., Malzbender, T.: A histogram-based color consistency test for voxel coloring. In: ICPR. (2002) 118–121

[12] Yang, R., Pollefeys, M., Welch, G.: Dealing with textureless regions and specular highlight: A progressive space carving scheme using a novel photo-consistency measure. In: ICCV. (2003) 576–584

[13] Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall (2002)

[14] Szeliski, R.: Rapid octree construction from image sequences. Computer Vision, Graphics and Image Processing: Image Understanding 58 (1993) 23–32

[15] Prock, A.C., Dyer, C.R.: Towards real-time voxel coloring. In: Image Understanding Workshop. (1998) 315–321



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY