International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

A Review of the Fundamentals of Free Viewpoint TV

Mohamad Raad, Majd Ghareeb, Ali Bazzi
Department of computer and communications engineering,
Lebanese International University
Beirut, Lebanon

***Abstract*— This paper reviews the concepts underlying Free View Point TV (FTV) technology. FTV promises to revolutionize the way in which we interact with captured visual scenes. Existing state of the art technology is based on concepts and theories that have been evolving over the past century. Nonetheless, major theoretic and conceptual hurdles still need to be overcome to allow the full potential of FTV to be reached. One of these is the ability to sample the visual scene accurately enough for accurate synthesis to be achievable. This paper presents the case for a fresh look at the way that visual scenes are sampled.**

Keywords—FTV, light field, free viewpoint

I. INTRODUCTION

Free View Point TV (FTV) promises to revolutionize the way we consume visual content [1]. FTV allows a viewer to see a visual scene from different points of view, either by moving around a display or by remaining stationary and selecting different points of view. In some way, FTV is an extension of 3D TV since 3D presentations rely on the rendering of two views to create a perception of depth by the viewer [2]. However, FTV is more than just an extension of 3D TV, it promises to allow a true immersive experience for the viewer who can interact with a complete visual scene rather than one view from that scene.

Some form of FTV has been used by broadcasters in limited trials [3], however, the choice of which view viewers saw in such trials was left up to the producer. The challenges that needed to be overcome in order to create limited versions of FTV include: the capture of the visual scene; the selection of the sub-set of possible views from that scene; the compression and transmission of the views selected; and the rendering of the selected views for the viewer.

The development of stereoscopic 3D display technology (requiring glasses) has been closely followed by the development of auto-stereoscopic display technology (which does not require glasses) [2, 4]. Such displays can be used to generate multiple views at different locations within the viewing region. It has recently been reported that displays capable of rendering more than 70 views have been developed [5]. With the evolution of the auto-stereoscopic display technology, the rendering of FTV content has become a feasible possibility.

This paper is a review of FTV and the concepts underpinning the technology enabling it. Although a whole system perspective is taken whereby the generation, representation, compression, delivery and rendering of FTV content is discussed, the main focus is on the models used to parameterize the problems that need to be solved. Recent reviews of FTV such as [1] and [5] have either provided a very general introduction to the topic (as in [1]) or a review that is very focused on one part of the FTV pipeline, namely compression via the use of depth maps (as in [5]). In contrast, this review aims to provide a survey of the concepts related to FTV.

A. *The Free View Point Challenge*

To allow immersion within a visual scene, that scene must first be captured. In order to capture a visual scene, one must know what to capture. As such, a conceptual model of a visual scene is required. One popular such model is the "Plenoptic function" [6] which represents the visual scene as a seven parameter function that is based on the idea that light can be represented by a set of straight lines or "rays" and these rays can be grouped into "pencils" via which a capturing device (such as a camera or an eye) gets a view of the scene. Thus the Plenoptic function is represented as:

$$F_P\left(\theta,\phi,\lambda,t,V_x,V_y,V_z\right) \qquad (1)$$

which represents the capturing device's position with the V terms, the angles at which the light rays enter the eye or camera with (θ,ϕ), the wavelength of the ray and the instant of time at which it is captured with (λ,t). This representation is typically simplified via the assumption of a constant wavelength and the dropping of the time term (because each image represents one sample in time) which leads to a five dimensional function. The five dimensional version of this function seems to have been initially discussed in the introduction to [7] where the translators of that work into English point out that Gershun's three dimensional vector representation is insufficient to represent the light field (which is another name for the Plenoptic function and will be used interchangeably with it). This five dimensional function is explained in [8] as follows[1]:

"*From an operational standpoint, the first basic photometric concept is radiant power per unit area. We measure this quantity at a point P in a definite plane. Since the position of a point in 3-space requires the specification of*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

three independent coordinates, and the orientation of a surface needs two angles, we are dealing with a scalar function of five variables. We propose to call this concept ***pharosage,*** *and to denote it by D. (Pharos is a Greek word meaning "lighthouse" or "beacon"; the age ending indicates "per unit area").*"

Having some mathematical representation for the visual scene then raises the question of how such a function should be sampled. In order to sample a function, re-parameterization may be required. It will be seen that this has actually been found to be the case for the light field. In turn, the frequency of sampling and the method of sampling become issues. Next, the representation of these samples for storage or transmission needs to be considered. These considerations must be made with the rendering and display tools in mind since these perform the necessary reconstruction of the visual scene from the, perhaps compressed, samples for presentation and navigation.

To the above list of issues, one may add the question of which part of the system does what? To clarify: for FTV to provide an immersive experience, it must recreate all of the views of the visual scene that the viewer wishes to view. Those views could be from outside of the scene, or inside the scene. Figure **II-1** illustrates the difference. To be able to recreate all those views, enough information must be available at the rendering point. If, as is expected, that information is to be compressed then there must be some way of interpolating between sample views (since compression, by definition, entails the discarding of redundant information). Such interpolation is computationally complex (the complexity of which will be elaborated upon in later sections) and hence the question of where those computations are to be spent arises.

B. Elements of a Free View Point System

Figure **II-2** illustrates the main components of a FTV system. The components shown within dashed boxes are those that are not essential for such a system but would probably be included in a practical FTV pipeline. From a system design perspective, the main components of FTV are scene capture, (potentially) scene editing, parameterization, compression, decompression, view interpolation, presentation and navigation control.

As will become apparent, although scene capture may be conceptually separate to parameterization, these two functions are interdependent (for example, specialized equipment may be used to enable a faster depth map generation, as will be discussed in section II.C). The same applies to compression and parameterization as well as view interpolation and parameterization. As such, a major part of this paper is spent on discussing the different parameterizations that have been used to represent visual scenes.

It will also become apparent that scene capture is not a straight forward task, involving resource trade-off decisions in terms of the number of views that will be captured and in what way (as in what camera technology will be used, what will be the visual scene boundary and over which period of time). This, of course, is scene sampling and it forms, in the author's opinion, a significant hurdle to the realization of FTV systems. We begin the next section with a discussion of sampling approaches that have been, and are, used for visual scene capture.

II. THEORY - FREE VIEW POINT MODELS

A. Sampling the light field

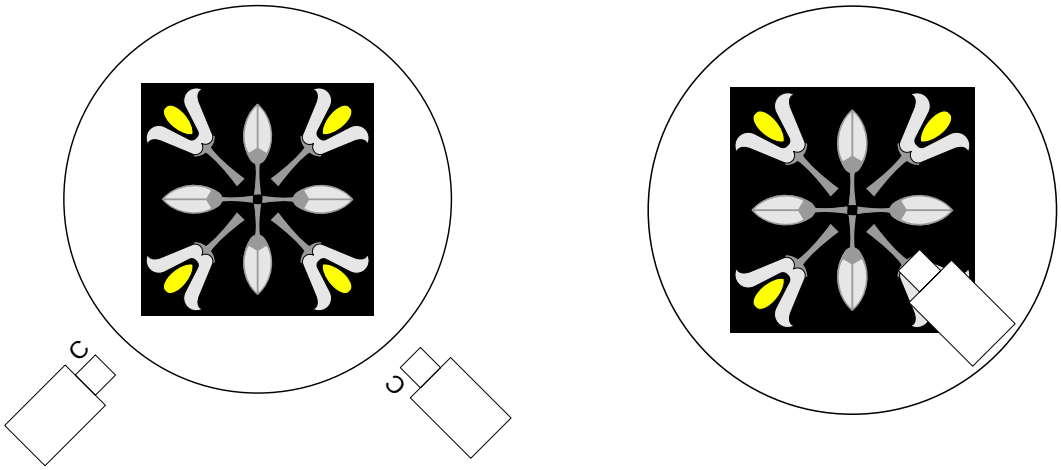
As mentioned, one of the first issues that arise in FTV is that of sampling the visual scene that FTV will allow the viewer to experience. According to sampling theory, a function may be represented by:

$$g(t) = \sum_n g(\lambda_n) S_n(t) \qquad (2)$$

when it is continuous in some domain [9]. In order to sample the light field, it is assumed that the five dimensional function (the reduced Plenoptic function) is continuous in all dimensions. Whether this assumption is absolutely correct, in a geometric sense, given what is known about the physical nature of light [10] is questionable. However, this is the assumption that has been made historically.

The light field is typically sampled with 2D images [11-14], so the placement and synchronization of cameras around the visual scene are significant considerations. In much of the early work on light field capture and reconstruction, specialized systems needed to be developed to provide what seemed to be a sufficient set of samples [12, 15, 16]. This situation has not changed much in recently reported work, although the focus has shifted to more challenging problems [17-19]. Earlier works in the field of Plenoptic function reconstruction focused on the use of images only as a step away from the then prevalent geometric representation of scenes [12, 14] because of the recognition of the complexity of obtaining accurate geometric representations of visual scenes [20, 21]. This raised the question of how many images were required to accurately represent a visual scene. Somewhat of an answer was provided in [11] where the authors identified an inverse relationship between the number of images and the complexity of geometric representation, where the complexity of the geometric representation refers to the number of unique depth maps available with the images. Now, the generation of accurate and consistent depth maps is not a trivial issue [5, 22-24] and so the trade-off between the number of images sampling the light field and the number of depth maps is in turn not trivial to optimize. However, the consensus in the field seems to be in favor of the use of depth maps in combination with image samples to represent a visual scene [5].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017



(a) (b)

Figure II-1 (a) views from outside the scene (b) a view from inside the scene

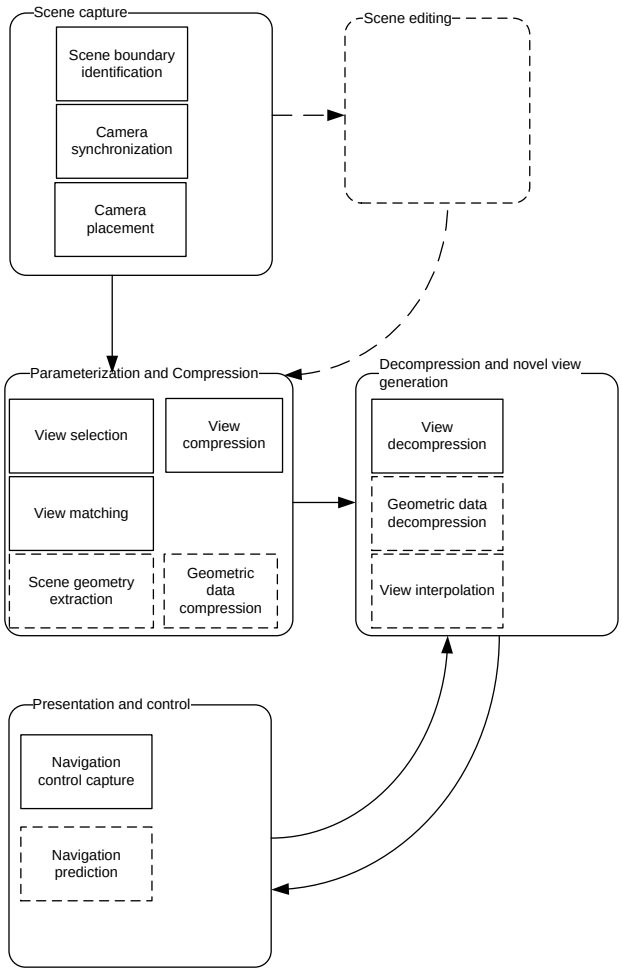


Figure II-2 High level decomposition of Free Viewpoint TV system

What the foregone discussion points to is a fundamental issue that does not seem to have been settled in the literature, specifically: what band-limiting should be applied to the signal representing the light field before sampling is undertaken? It is well known from sampling theory that the general assumption underlying the representation of a function as given in (2) is that it is band-limited. Thus, in order to sample a function sufficiently, the bandwidth of that function needs to be known (or imposed on the original function through filtering). This does not seem to have been taken into consideration in the previously mentioned works. If one considers the sampling setup described in [25] where the authors describe the capturing of a "ground truth" image set to which interpolated views can be compared, it is noteworthy that the number of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

sampling cameras is considerably more than those in previously reported works. Yet the authors do not claim that their capture setup is sufficient to capture all of the non-redundant visual information of the light field (nor do they need to as they were concerned with the results of algorithms interpolating between views – to view a complete result set for the tested algorithms see [26]). The point here is simply that it may be impractical to sample a light field completely in the traditional, band limited, sense (how many images would need to be captured in a second to ensure that absolutely no changes in a view were missed?). Instead, sampling the light field may require the use of sampling techniques for "not necessarily band-limited functions" [27].

B. Light Field Representation

The Plenoptic function represents the light field in one set of parameters, others have been (and continue) to be used. The main reason for the use of different parameter sets is because this is one way of restating the (mathematical) problem at hand to allow the application of proven tools for the development of a solution. Summaries of popular parameterizations are presented in [20, 29]. Following the nomenclature of [29], these parameterizations can be labelled as the "two plane model" (2PP), the "point on a plane with direction model" (DPP), the "two points on a sphere model" (2SP) and the "point and direction model" (PDP)

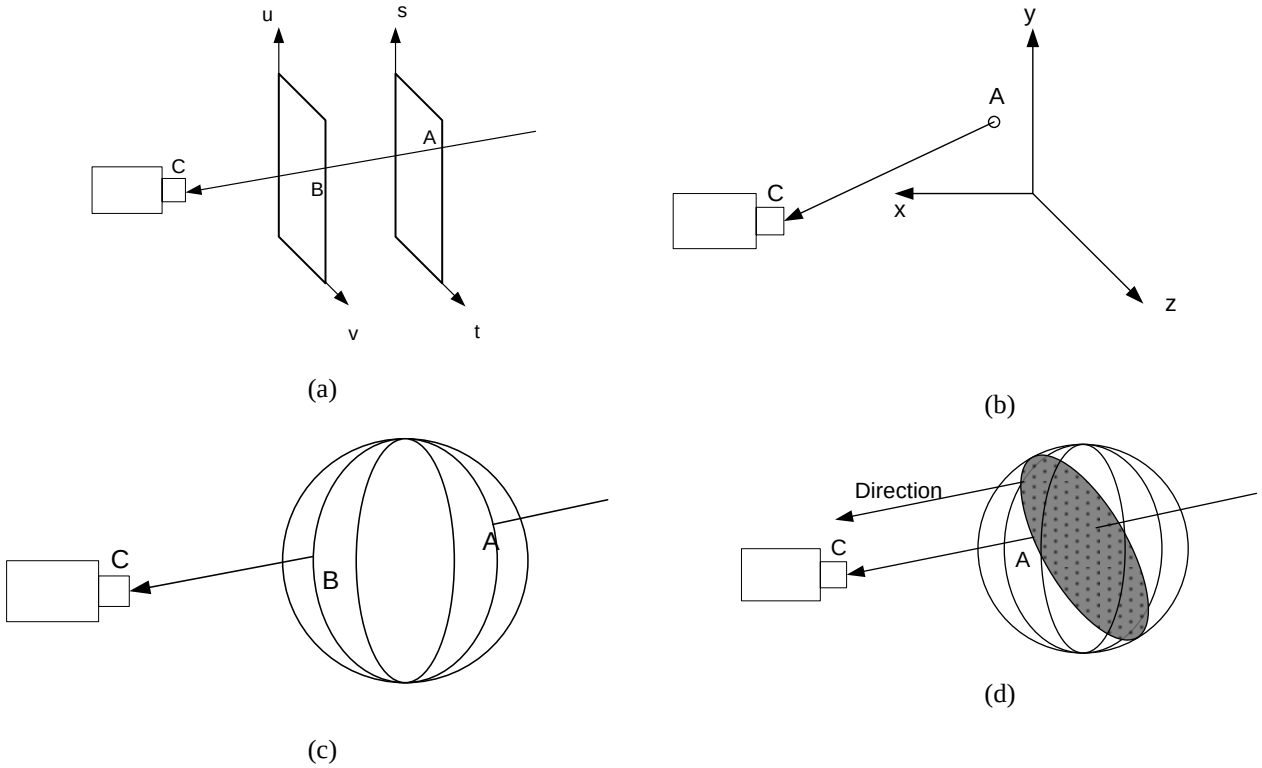


Figure I-1 (a) two plane parameterization (2PP), (b) point and direction (PDP), (c) two points on a sphere (2SP) and (d) direction and point (DPP)

To highlight the significance of obtaining a deeper understanding of the sampling problem when it comes to the light field, consider the discussion presented in [13] with regards to the potential applications of light field rendering. In focusing on one potential application, the author of [13] discusses how increased sampling density leads to not only the ability to reconstruct views from the same plane as the capturing cameras, but also to reconstruct views from a plane that is closer to the object (allowing a zoom in of the light field, as mentioned previously). It is further pointed out that one of the main challenges at the time of writing was the capture, parameterization and reconstruction of a reflectance field – i.e. a light field that does not assume constant illumination, but rather a combination of incoming rays and reflected rays. This is clearly the more realistic representation of our world and the influence of reflectance on an entire light field seems to have been considered first in [28].

1) Light Field – DPP (Ray Space)

The Ray Space parameterization (which is the same as DPP) has been advocated by one of the pioneers of FTV [1, 30] because it aids in the conceptual design of practical systems. However, another advantage claimed for this parameterization is that it enables uniform sampling because it fits the spherical model of the light field according to [29], whose analysis of different light field models results in an error bound metric for the different parameterizations. According to this metric, all parameterizations that fit the spherical model of the light field have lower direction and position error bounds (these are the components of the developed error bound metric) because of the ability to uniformly sample the sphere. However, these are the error bounds of the available models and not the error incurred due

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

to a signal processing function (such as quantization) and these bounds are derived fundamentally based on the number of positional and directional samples.

One interesting extension to such an analysis would be to determine the sensitivity of each model to signal processing noise, which would be of relevance to FTV because of the expected production chain (capture, parameterization, compression, transmission/storage, decompression and rendering). Further, it would also be relevant to carry out a similar analysis with a different sampling series to that used in [29] to gain an understanding of whether the above mentioned conclusion is a manifestation of the parameterization used or of the sampling series assumed.

The use of DPP has also been advocated in [23] where the concept of sprites with depth and layered depth images was introduced. The sprites with depth and layered depth images allow for smooth interpolation between views. The layered depth maps store multiple depth values for a single line of sight, allowing for faster rendering of different views. Although the main impetus for these concepts at that time was the complexity of rendering, the authors present a relevant discussion about the sampling of the light field as represented in DPP. The authors argue for a sampling arrangement where only rays carrying significantly different information from what is available are accepted as new samples. Naturally, one would need to identify the light rays which carry "new" information and that is possible if the scene geometry is known along with the capture setup. More discussion on the use of depth maps in combination with images will be presented in a later section of this paper.

In a significant shift towards enabling high quality view reconstruction, [16] reported the innovative step of representing the 4D DPP of the light field as a summation of the products of 2D functions, as given by (3) [16].

$$f(r,s,\theta,\phi) \approx \sum_{k=1}^{K} g_k(r,s) h_k(\theta,\phi) \qquad (3)$$

The first 2D function represents the surface light field – i.e. the light field at the source, and the second represents the view from which the surface is being observed. This separation clearly reduces the need to make the simplifying assumption that the surface has constant illumination as the original light field parameterizations did (such as in [12]). This "factorization" of the light field also allows for very high compression rates to be achieved (1000:1 and more were reported in [16]) whilst maintaining high PSNR measured quality and the ability to render on standard graphics cards in real time (with the image size used by the authors). The reason for the high compression ratio obtained is the ability to apply different compression techniques to the two functions. It is noteworthy that the authors do not explain the reasons for their choice of factorization, but its power is clearly in separating the source from the recorded view, allowing for mathematical associations between different views of the same source. The system presented by the authors is an image based capture and rendering system that requires a specialized image capture setting, as did almost all other visual scene recreation systems at that time. Other works have since extended this concept of factorizing the visual scene such as that presented in [31] where the authors of that work propose a factoring of images which is applicable to both rendering and compression.

2) Light Field - 2PP

One of the earliest and most widely cited 2PP representations is described in [12] where a three dimensional cube is described as six flat plane pairs (named "slabs") in order to capture the different views and to enable the generation of new views. In this case, the Plenoptic function is reduced to a 4D function and then parameterized using two planes per view. The distance between the planes is fixed. A special stage with visual tags was used by the authors to allow the determination of the visual cube. It is significant to note that the authors used a quadralinear basis function for their sampling because it is not clearly explained why that choice had actually been made. It would have seemed more in line with existing sampling theory to actually use a three dimensional sinc function (or an appropriate version of the sinc function) [9]. It seems that the use of a quadralinear function could be the reason why the authors then need to use depth estimates in order to avoid blurring of the image. This approach is very similar to the one described in [5].

Further, the authors of [12] do not address what happens at the edges of the cube – which would clarify how this approach would compare to spherical parameterization. One could guess that the authors would allocate the edge between two slabs (a pair of planes) to either, however, the rays from the slabs would cross over at the edges which would most likely lead to a "jump" in between views. One of the main problems that [12] faces is the lack of an error or distortion function that is being minimized. So the results presented are in terms of practical implementation and some visual evidence showing that the presented method works.

These observations of the limitations of the 2PP of the light field were also made in one of the earliest models of the light filed as described in [32]. The model developed in that work is cylindrical because of the recognized disadvantages of using a cubic model (the oversampling problem between the faces of the cube, as noted above) and the spherical models (it was difficult to capture the information in that format for programming purposes at that time). Although the authors recognize that the cylindrical model is in fact an approximation of the more complete spherical model, there is no mention of any assumptions regarding the properties of the surfaces that are being represented in the views which make up the Plenoptic function. The authors use a transformational relationship between images (any two images from different views can be shown to be transformed versions of one another), which has since become a popular part of image based rendering, where the transform is decomposed into intrinsic (internal to the camera) and extrinsic (external to the camera) properties, however, the surface attributes of the objects being represented by the images are not captured by this model.

3) Free Viewpoint TV and different light field representations

A primary requirement for FTV is the ability to interpolate new (or novel) views from existing views (samples). As such, it is worthwhile for us to take a deeper look into how different

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

representations of the light field can influence the results of such an operation. In this section it is assumed that the light field representations do not utilize scene geometry information, the use of which will be discussed in section II.B.4).

According to [32], "all image-based rendering approaches can be cast as attempts to reconstruct the plenoptic function from a sample set of that function"[2] and that "the most natural surface for projecting a complete plenoptic sample is a unit sphere centred about the viewing position."[3] However, the authors then go on to claim that at the time they were developing their system there was no representation that could capture such a projection on a computer. As such the authors identify the superiority of the spherical representations discussed previously over 2PP and the practical difficulties of those representations. Because of these practical difficulties, a cylindrical representation is then developed in [32] which can be thought of as a "half way" solution between the practical, but inaccurate 2PP representation and the (at that time) impractical, but conceptually correct spherical representation.

In contrast, the approach described in [33] (which introduced both 2SP and DPP) solves the problem of spherical sample storage by fitting the sphere to be sampled tightly around an object in the visual scene and tracing rays in the 3D space in the direction of the eye or camera (the required view) to determine where these rays cross this sphere (if they do at all). Note that in this case the sampling is carried out for a sphere approximating the convex hull of the object whereas the discussion presented in [32] considered fitting the sphere around the view position. This explains the ability of the authors of [33] to define a data structure which represents either 2SP data or DPP data. 2SP data is represented via the identification of the two triangles through which a ray passes (the sphere is generated from icosahedron triangular projections) whereas DPP data is represented through the intersection of the ray with a great circle of the sphere and the triangle on the surface of the sphere through which the ray passes (as illustrated in Figure -3 (d)).

Since the spherical based schemes actually use approximations of a sphere, absolute uniform sampling is only approximately achieved. Yet, in comparison with 2PP the reported retrieval and rendering functions are significantly more complex (2-3 times more complex according to [33]). In return for this increase in complexity, the disparity problem (the introduction of artefacts as viewing position is moved between slabs in the 2PP model) is almost completely eliminated (we say "almost" here because the sphere being sampled is only approximated). Whether or not this return is worth the increased complexity cost depends on the novel (interpolated) views one expects to generate, and the computational resources available. It is expected that practical FTV systems will not be constrained in terms of computational resources, so if one wishes to provide true immersion in a visual scene then one of the spherical representations should be adopted if the camera technology being used allows for the capture of images that lead to the DPP or 2SP parameters being extracted (in the case of 2SP ray tracing is a requirement of the capturing system). Such a system is discussed in [1] but that system is most suitable for DPP. It is notable that the main advantage that the spherical based representations over a cylindrical representation is that of constant quality regardless of view direction (as in the visual scene can be circumscribed in terms of views in all directions). Finally, DPP also allows for the use of depth maps in association with the available view samples to enhance the quality of the rendered views. The combination of scene geometry with image samples will be discussed next.

4) Scene Geometry and Image Based Rendering

The approach taken by [32] sits on a grey border separating visual scene reconstruction through purely light field rendering from available image samples and scene reconstruction that uses both geometric representation and image based rendering. It has been long acknowledged that the combination of scene geometry with Image Based Rendering (IBR) is probably the most practical approach to light field reconstruction available [13]. Image based rendering is a vast field that has seen application in, and contributions from, the fields of robotics, signal processing, computer vision and photogrammetry [34]. Although we will not delve into the details of all of the techniques that have been developed in this field, it is important to recognize the relationship between image based rendering and FTV.

The aim of IBR, as described in the pioneering work of [32], can be stated as: "*Given a set of discrete samples (complete or incomplete) from the plenoptic function, the goal of image-based rendering is to generate a continuous representation of that function.*"[4]. in other words, the focus of IBR is on the extraction of "novel" views from available views. Now, it is not necessarily the case that FTV requires IBR in that it is conceivable that an FTV system would have enough actual views available to it to allow perceptually seamless viewer navigation. However, this is highly unlikely due to the volume of information that one would need to create such a representation of a visual scene. The more likely scenario is that only a subset of views will be available and that subset will be used to generate other views to allow viewer navigation of the scene [5, 13, 21]. This is recognized, for example, by the Moving Pictures Experts Group (MPEG) to be the most likely scenario and that is why depth map extraction technology and "view synthesis" (including interpolation) technology has been sought in MPEG's call for proposals on 3DV [35] (which is an international standardization effort that is part of MPEG's push to standardize FTV representation technologies).

The difficulty of generating the geometric representation of a scene (for representation as depth maps, for example) has been extensively discussed in the literature and interested readers are referred to works such as [36], who discuss a 3D model generation in a real time system, for further details. It is important to appreciate the difficulty of generating such models to contextualize the practical relationships between sampling, parameterization and viewer experience. Such geometric models are required when one decides to use a system that needs scene geometry in order to render novel views. This choice influences the sampling required as well as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

the parameterization used. It also adds the cost of generating the geometric models.

A different approach (one based on Photogrammetry techniques) was taken in [37] where a number of image matching methods for the extraction of geometric scene information from images are presented. The authors report the achievement of fairly accurate depth reconstruction from images of the same scene. Although it is not clear if consensus exists in the field regarding the metric used to determine accuracy (especially when no reference is available). The authors apply their approach to the 2PP parameterization of the light field and report good quality rendering results in real time. However, how quality was measured by the authors (besides visual inspection) is not clear. The issue of developing objective quality measures for interpolated views will be returned to later in this paper.

One of the earliest practical FTV systems that employs a combination of image based rendering with geometric information (depth maps) is presented in [21]. That system allows for the generation of new views from existing views through view blending. The core of the system relies on the separation of "boundary depth" (around the edges of objects) layer from a main layer and blending these separately. The reason behind that work was extending IBR to dynamic scenes, where the synchronization of many cameras becomes an issue (as does the large number of images that are required to perform IBR when scene geometry is unknown). The authors used "matting" to reduce depth discontinuities. One may consider this as a type of interpolation of the scene geometry, allowing for the reduction in image samples to create a smooth flow between views – in line with the ideas presented in [11]. Still, the authors need to use novel hardware for their scene capture and the use of a specialized codec highlights the intimate relationship between scene capture, parameterization and compression (where the first two have a major influence on the design of the third).

In an effort to categorize the various approaches taken to capture the geometry of a visual scene, [25] breaks down scene representation approaches into three tracks of investigation:

1. Geometry on a 3D grid:
    a. Those algorithms that use voxels (a value on a 3D grid representing a 3D picture element, or the equivalent of a pixel in 3D [38]), such as [39], [40], [41], [42], [43], [44] and [45].
    b. Those that use "encoding distance to the closest surface" such as [46], [47], [48] and [49].
2. Polygon meshes [50], as used in [51] and [52].
3. Depth map representation, as used in [24], [21], [53], [22] and partially surveyed in [5]. A complimentary approach is the use of "relief fields" instead of depth, i.e. the use of height above an approximated geometric representation as described in [54].

Again, the purpose of most of these methods is to enhance the quality of the interpolated views (whilst some focus on maintaining quality for a reduced computational complexity) through a combination of available views with more accurate geometric representation of the visual scene. The effort expended towards meeting this goal by researchers is testament to the difficulty of reaching an optimal balance between view samples (images) and geometric representation, as discussed earlier.

It should be clarified that the view generation algorithms themselves are different to the scene geometry modelling components, and these have also been classified by [25] into roughly four categories:

1. Those that extract a surface from a 3D volume upon which a cost function has been computed (as in, create a 3D volume and then take a view of it);
2. Algorithms that tinimize some cost function through iteration (with the aim of moving closer to what is believed to be the actual view of a volume);
3. Those that compute depth maps and ensure that the depth maps are consistent with each other according to a cost or error function; and
4. Algorithms that fit reconstructed surfaces to a set of extracted feature points as in [55], [56] and [57], which basically reduces the focus of the reconstruction algorithm to be on those feature points.

Recently, [58]and [17] have reported methods that deal with the more challenging problem of view reconstruction from handheld and moving cameras, a much more challenging problem than has been dealt with in most of the visual scene reconstruction literature. It is argued in [17] that image based rendering with the explicit use of geometry is more appropriate for applications such as FTV than purely light field and implicit geometry rendering because of the success that such approaches have had in the past. Thus, the approach advocated is an image rendering approach that uses explicit geometry (this term refers to algorithms that require a geometric model of the scene, such as a depth map, and seems to have been first introduced in [34]) and is an extension of that presented in [59]. The method presented in [17] relies on multi-layered segmentation of the different images sampling the visual scene. The results presented are claimed to be on par with [60] which achieves sub-millimetre accuracy in view interpolation compared to ground-truth images (according to results provided on [26]). However, this objective measure does not seem to capture the distortion that is visually apparent from that algorithm. Nonetheless, those papers report significant improvements in view interpolation with very promising results for application in the FTV space, and it is significant to note that both rely on an evolved combination of scene geometry and IBR.

Prior to the above reported works, a slight paradigm shift had occurred in research into visual scene reconstruction using IBR and geometric modelling, specifically the move to separate a "visual hull" from the scene and to deal with that hull separately to the objects "embedded" within it, as in [18, 61]. Both of these systems utilise segmentation and the visual hull method to render virtual views. The computational demands of such an approach are discussed in [18] and it is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

clear that such an approach could only operate off-line at the time it was developed. There were also errors reported with some of the virtual views generated (for example some views would have missing players – or moving objects). The latter system is somewhat an extension of the ideas discussed in [19] where the objective there was the generation of virtual views on a tennis court (naturally, considerable interest in the application of FTV comes from sports content distribution). The approach taken there was also to segment the available images, to project the segments towards virtual views and then to recombine the projected segments to form a new virtual view. Although interesting, and of potential use, from an FTV perspective, it is not clear whether the reported works have resulted in practically useful systems given the lack of objective results reported in terms of both quality and complexity in the literature to date. However, recent work focusing on the development of numerical solutions for the geometric re-projection of images (and therefore enhancing the practicality of such an approach) such as [62, 63] suggests that this type of interpolation by segmentation approach could become practical in the near term. If that were to occur, the quality of the results reported in the previously discussed works suggests that this approach will become much more widely adopted.

C. The different representations and Free Viewpoint TV

The described methods of modelling a visual scene (geometric or light field) are not exclusive, but they do lead to different practical systems. The use of depth maps in the interpolation of novel views has necessitated the development of specialized cameras, such as that described in [64] which detects scene depths through the use of infrared light analysed using the camera shutter speed. The use of the light field model has also lead to the development of specialized cameras which are succinctly described in [65] that use microlenses, positioned behind the main camera capturing lens to separate rays that have been focused by the main lens, so that different views can be accessed – or different rays arriving at the same view position can be accessed. These can then be used to generate a depth map of the scene also, using a method such as that described in [11].

Given the above, either model could be used for the generation of FTV content. However, the use of depth maps allows for a reduction in the number of views (images) required and it provides some way of "linking" actual views with novel (interpolated) views. So the use of geometric information (in the form of depth maps) in combination with image based rendering seems the best approach for the creation of FTV content. Yet, even with this approach it is essential to determine how many actual views are required to allow immersive navigation of the visual scene. For that, the light field model provides some guidance by identifying the dimensions along which the scene must be sampled. The sampled light field will be a representation of the visual scene "band-limited" along each of these dimensions. The band-limiting required will depend on the scene (the source) being sampled (as does band-limiting in other multimedia functions such as sound capture). Thus, knowledge of the scene (not just geometric) will be required for either of the discussed models to be supplied with sufficient information to allow for high quality reconstruction. Section III further discusses such issues.

III. CONCEPTUAL AND PRACTICAL CONSIDERATIONS

Although practical systems allowing for visual scene reconstruction and navigation have existed for some time [66], gaps remain in the visual scene capture, representation and reconstruction body of knowledge. One such gap that exists in the current state of the art is for an objective quality measure of interpolated (synthetic) views. Some attempts have been made to develop such a quality measure, for example in [67, 68]. The measure described in [68] is based on the L_2 distance that a given portion (say *M%* of all the pixels in an image) of a generated virtual view is from a ground truth sample (image) from that view. This means that the measure can be used to provide some indication of quality if the reference is known. It is claimed that this measure can provide a no reference quality indicator as well, although that is simply a comparison between the re-projections of existing cameras. The developed metric uses optical flow [69] to locate similar pixels in different images. Clearly, this measure is content dependent but it can be used to provide comparative results between algorithms or systems. In contrast, the metric reported in [67] operates purely on resultant images, however, the presented subjective results do not correlate well with the outcomes of the proposed metric. Still, it seems that such a metric, with its focus on the ghosting artefact (which is visually prominent in interpolated views) does provide some comparative indication of quality (as in it could be one metric in a basket of metrics that two systems or algorithms can be compared with).

A prominent and practical approach that has been taken to evaluate view interpolation algorithms is described in [25], and continues to be updated via [26]. The authors of [25] developed a performance framework based on the error between a reconstructed surface and a "ground truth" surface of the same object. The ground truth surface is generated through a highly sampled hemisphere around the object using calibrated cameras. The results are reported in terms of how far away 90% of the points on the reconstructed surface are from the equivalent points on the ground truth surface as well as on how many of the ground truth points have an equivalent in the reconstructed surface (as in how much of the view has been actually reconstructed). Whilst these measures seem conceptually very indicative of quality, it is clear upon examining the visual results shown on [26] that there is a significant difference between the results based on these measures and visual similarity. In fact, the results reported are not consistent with expected performance, for example some algorithms reported perform "better" with sparser samples of the light-field around the target object (fewer images) according to these measures.

The above two observations leads one to conclude that there is a need to develop *perceptually accurate* reconstruction algorithms rather than *mathematically accurate* representations (a review of what is known about 3D perception, or stereo vision, and what attributes a good perceptual quality presentation should have is presented in [70]). The work reported in [71] presents such a paradigm shift where the focus is on image interpolation that is perceptually smooth rather than mathematically accurate (as much of the previous work has been biased towards). That work is an extension of [72] where view interpolation is approached without the need for the reconstruction of 3D geometry. Also of relevance to this area are works such as [73] which focus on image interpolation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

that produces a "convincing" result rather than accurate results. These, and similar, developments raise interesting questions about services such as FTV which could, conceivably, present visually pleasant results, but not scene accurate ones. It is highly unlikely that such an approach would become widespread in a FTV service simply because of the need for temporal continuity (as in, although the image interpolation results may be pleasant per image, that may not be the case for a video sequence).

The same approach as in [25] is currently being taken towards optical flow tracking algorithms [74] and interested parties can still submit their algorithms for evaluation via [26]. Optical flow, as developed originally in [69], has been applied mostly in robot vision for view interpolation and may find some application in FTV, however this seems unlikely given the restricting nature of the assumptions upon which optical flow algorithms are based on (specifically the constraining assumption that surfaces have constant luminance). In any case, the conceptual underpinnings of optical flow are very like those used in video compression for motion compensation and thus could still see application in FTV.

IV. CONCLUSION, CHALLENGES AND FUTURE DIRECTION

This paper has presented a survey of the main concepts underpinning the technologies upon which Free viewpoint TV depends. These concepts deal with the capture and sampling of a visual scene (regularly referred to as the light field), the parameterization of that scene for information representation, the compression of the extracted information and the interpolation of novel views from sampled views.

A number of conclusions may be drawn from this survey. First, there is a need for a clearer association of light field capture with sampling theory. The surveyed attempts at explaining light field sampling do not provide a clear association whereby practical decisions can be inferred regarding the sufficiency of information available for high quality (or perhaps lossless) reconstruction. There is also a need for an objective quality measure that can be used to determine interpolated image quality. If such a measure were to have a "no reference" version, a major contribution would be made towards enabling the deployment of FTV systems. Recent attempts at developing such a measure are encouraging but still fall short of filling the existing gap. Further, the recent trend towards using samples that have been captured from mobile or handheld cameras accentuates the necessity to fill these gaps. Finally, it is also important to quantify the computational requirements of FTV as a complete system in order to identify bottlenecks that should become the focal points for further investigation.

Looking towards the future, as FTV becomes an accessible reality to multimedia consumers, demand will grow for it to be provided over heterogeneous networks and devices. This will necessitate increased efficiency in terms of visual scene reconstruction in order to maintain a high quality, visually immersive experience. This, in turn, could (and, in the author's opinion, probably will) lead to new compression and rendering technologies that will both need a solid theoretical basis through which visual scenes can be better understood. As such, we expect Free viewpoint TV to witness significant technological leaps in the next three to five years.

V. REFERENCES

[1] M. Tanimoto, et al., "Free-Viewpoint TV," IEEE Signal Processing Magazine,, vol. 28, pp. 67-76, 2011.
[2] N. A. Dodgson, "Autostereoscopic 3D Displays," Computer, vol. 38, pp. 31-36, 2005.
[3] T. Kanade and P. J. Narayanan, "Virtualized Reality: Perspectives on 4D Digitization of Dynamic Events," Computer Graphics and Applications, IEEE, vol. 27, pp. 32-40, 2007.
[4] N. S. Holliman, et al., "Three-Dimensional Displays: A Review and Applications Analysis," IEEE Transactions on Broadcasting,, vol. 57, pp. 362-371, 2011.
[5] K. Muller, et al., "3-D Video Representation Using Depth Maps," Proceedings of the IEEE, vol. 99, pp. 643-656, 2011.
[6] E. H. Adelson and J. R. Bergen, "The Plenoptic function and the elements of early vision," in Computational models of visual processing, M. Landy and J. A. Movshon, Eds., ed MA: MIT Press, 1991, pp. 3-20.
[7] A. Gershun, "The light field (Moscow 1936. Translated by P. Moon and G. Timoshenko)," Journal of Mathematics and Physics, vol. 18, pp. 51-151, 1939.
[8] P. Moon and D. E. Spencer, The Photic Field. Cambridge, Massachusetts: The MIT Press, 1981.
[9] J. R. Higgins, Sampling Theory in Fourier and Signal Analysis. Oxford, UK: Clarendon press, 1996.
[10] H. Haken, Light vol. 1. Amsterdam: North-Holland Publishing Company, 1981.
[11] J.-X. Chai, et al., "Plenoptic sampling," presented at the Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000.
[12] S. J. Gortler, et al., "The lumigraph," presented at the Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996.
[13] M. Levoy, "Light Fields and Computational Imaging," Computer, vol. 39, pp. 46-55, 2006.
[14] M. Levoy and P. Hanrahan, "Light field rendering," presented at the Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996.
[15] T. Kanade and P. J. Narayanan, "Virtualized Reality: Perspectives on 4D Digitization of Dynamic Events," IEEE Journal of Computer Graphics and Applications, vol. 27, pp. 32-40, 2007.
[16] W.-C. Chen, et al., "Light field mapping: efficient representation and hardware rendering of surface light fields," ACM Trans. Graph., vol. 21, pp. 447-456, 2002.
[17] J.-Y. Guillemaut and A. Hilton, "Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications," International Journal of Computer Vision, vol. 93, pp. 73-100, 2011.
[18] N. Inamoto and H. Saito, "Virtual Viewpoint Replay for a Soccer Match by View Interpolation From Multiple Cameras," IEEE Transactions on Multimedia, vol. 9, pp. 1155-1166, 2007.
[19] K. Kimura and H. Saito, "Player viewpoint video synthesis using multiple cameras," in Proceedings of the 2nd IEE European Conference on Visual Media Production, 2005 (CVMP 2005). , 2005, pp. 112-121.
[20] E. Camahort, "4D light-field modeling and rendering," PhD Thesis, Faculty of the graduate school, University of Texas at Austin, Austin, 2001.
[21] C. L. Zitnick, et al., "High-quality video view interpolation using a layered representation," ACM Trans. Graph., vol. 23, pp. 600-608, 2004.
[22] P. Gargallo and P. Sturm, "Bayesian 3D modeling from images using multiple depth maps," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005). , 2005, pp. 885-891 vol. 2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

[23] J. Shade, et al., "Layered depth images," presented at the Proceedings of the 25th annual conference on Computer graphics and interactive techniques, 1998.
[24] R. Szeliski, "A multi-view approach to motion and stereo," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999.. 1999, p. 163 Vol. 1.
[25] S. M. Seitz, et al., "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2006, pp. 519-528.
[26] D. Scharstein. (2011). Multi-view stereo evaluation web page. Available: http://vision.middlebury.edu
[27] P. L. Butzer and R. L. Stens, "Sampling theory for not necessarily band-limited functions: a historical overview," SIAM Review, vol. 34, pp. 40-53, March 1992.
[28] A. S. McAllister, Graphical solution of problems involving plane surface lighting sources: The law of conservation as applied to illumination calculations. The absorption-of-light method of calculating illumination. The bearing of reflection on illumination. Reprints of articles and paper. New York, 1912.
[29] E. Camahort, et al., "A line-space analysis of light-field representations," Graphical Models, vol. 71, pp. 169-183, 2009.
[30] M. Tanimoto, "Overview of free viewpoint television," Signal Processing: Image Communication, vol. 21, pp. 454-461, 2006.
[31] H. Wang, et al., "Factoring repeated content within and among images," ACM Trans. Graph., vol. 27, pp. 1-10, 2008.
[32] L. McMillan and G. Bishop, "Plenoptic modeling: an image-based rendering system," presented at the Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, 1995.
[33] E. Camahort, et al., "Uniformly Sampled Light Fields," presented at the Proceedings of the Ninth Eurographics Workshop on Rendering, Vienna, 1998.
[34] S. B. Kang, et al., "Image-Based Rendering," Foundations & Trends in Computer Graphics & Vision, vol. 2, pp. 173-258, 2006.
[35] MPEG, "Call for Proposals on 3D Video Coding Technology (w12036)," Video/Requirements, Ed., ed. Geneva, Switzerland: ISO/IEC JTC1/SC29/WG11, 2011.
[36] S. Rusinkiewicz, et al., "Real-time 3D model acquisition," ACM Trans. Graph., vol. 21, pp. 438-446, 2002.
[37] M. Pollefeys, et al., "Visual Modeling with a Hand-Held Camera," International Journal of Computer Vision, vol. 59, pp. 207-232, 2004.
[38] S. Dzik and J. Ezrielev, "Representing surfaces with voxels," Computers & Graphics, vol. 16, pp. 295-301, 1992.
[39] P. Eisert, et al., "Multi-hypothesis, volumetric reconstruction of 3-D objects from multiple calibrated camera views," in Proceedings of the IEEE International Conference onAcoustics, Speech, and Signal Processing, 1999. , 1999, pp. 3509-3512 vol.6.
[40] K. N. Kutulakos and S. M. Seitz, "A Theory of Shape by Space Carving," International Journal of Computer Vision, vol. 38, pp. 199-218, 2000.
[41] K. Kutulakos, "Approximate N-View Stereo Computer Vision " in Proceedings of ECCV 2000. vol. 1842, ed: Springer Berlin / Heidelberg, 2000, pp. 67-83.
[42] R. Bhotika, et al., "A Probabilistic Theory of Occupancy and Emptiness, Computer Vision — ECCV 2002." vol. 2352, A. Heyden, et al., Eds., ed: Springer Berlin / Heidelberg, 2002, pp. 112-130.
[43] Y. Ruigang, et al., "Dealing with textureless regions and specular highlights - a progressive space carving scheme using a novel photo-consistency measure," in Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003. , 2003, pp. 576-584 vol.1.
[44] G. G. Slabaugh, et al., "Methods for Volumetric Reconstruction of Visual Scenes," International Journal of Computer Vision, vol. 57, pp. 179-199, 2004.
[45] G. Vogiatzis, et al., "Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 2241-2246, 2007.
[46] O. Faugeras and R. Keriven, "Variational principles, surface evolution, PDEs, level set methods, and the stereo problem," IEEE Transactions on Image Processing,, vol. 7, pp. 336-344, 1998.
[47] J. P. Pons, et al., "Variational stereovision and 3D scene flow estimation with statistical similarity measures," in Proceedings of the Ninth IEEE International Conference onComputer Vision, 2003., 2003, pp. 597-602 vol.1.
[48] H. Jin, et al., "Multi-view stereo beyond Lambert," in Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, pp. I-171-I-178 vol.1.
[49] J. P. Pons, et al., "Modelling dynamic scenes by registering multi-view image sequences," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005). 2005, pp. 822-827 vol. 2.
[50] M. Martti, "Topological analysis of polygon meshes," Computer-Aided Design, vol. 15, pp. 228-234, 1983.
[51] A. P. Rockwood and J. Winget, "Three-dimensional object reconstruction from two-dimensional images," Computer-Aided Design, vol. 29, pp. 279-285, 1997.
[52] C. Hernández Esteban and F. Schmitt, "Silhouette and stereo fusion for 3D object modeling," Computer Vision and Image Understanding, vol. 96, pp. 367-392, 2004.
[53] V. Kolmogorov and R. Zabih, "Multi-camera Scene Reconstruction via Graph Cuts," in Proceedings of Computer Vision - ECCV 2002. vol. 2352, A. Heyden, et al., Eds., ed: Springer Berlin / Heidelberg, 2002, pp. 8-40.
[54] G. Vogiatzis, et al., "Reconstructing relief surfaces," Image and Vision Computing, vol. 26, pp. 397-404, 2008.
[55] O. D. Faugeras, et al., "Representing stereo data with the Delaunay triangulation," Artificial Intelligence, vol. 44, pp. 41-87, 1990.
[56] A. Manessis, et al., "Reconstruction of scene models from sparse 3D structure," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000. , 2000, pp. 666-671 vol.2.
[57] C. J. Taylor, "Surface reconstruction from feature based stereo," in Proceedings of the Ninth IEEE International Conference onComputer Vision, 2003., 2003, pp. 184-190 vol.1.
[58] L. Ballan, et al., "Unstructured video-based rendering: interactive exploration of casually captured videos," ACM Trans. Graph., vol. 29, pp. 1-11, 2010.
[59] J. Y. Guillemaut, et al., "Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes," in Proceedings of the IEEE 12th International Conference on Computer Vision, 2009 2009, pp. 809-816.
[60] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1362-1376, 2010.
[61] O. Grau, et al., "3D modelling and rendering of studio and sport scenes for TV applications," presented at the Proceedings of WIAMIS, Montreux, Switzerland, 2005.
[62] S. Agarwal, et al., "Fast algorithms for L_∞ problems in multiview geometry," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008 (CVPR 2008). 2008, pp. 1-8.
[63] F. Lu and R. Hartley, "A Fast Optimal Algorithm for L_2 Triangulation," in Computer Vision – Proceedings of ACCV

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

2007. vol. 4844, Y. Yagi, et al., Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 279-288.
[64] M. Kawakita, et al., "High-definition real-time depth-mapping TV camera: HDTV Axi-Vision Camera," Opt. Express, vol. 12, pp. 2781-2794, 2004.
[65] T. G. Georgiev, "Plenoptic camera with large depth of field," USA Patent, 2011.
[66] S. E. Chen, "QuickTime VR: an image-based approach to virtual environment navigation," presented at the Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, 1995.
[67] K. Berger, et al., "A ghosting artifact detector for interpolated image quality assessment," in Proceedings of the IEEE 14th International Symposium on Consumer Electronics (ISCE), 2010 2010, pp. 1-6.
[68] J. Starch, et al., "Objective Quality Assessment in Free-Viewpoint Video Production," in Proceedings of the 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008, 2008, pp. 225-228.
[69] B. K. P. Horn and B. G. Schunck, "Determining optical flow," Artificial Intelligence, vol. 17, pp. 185-203, 1981.
[70] G. Westheimer, "Three-dimensional displays and stereo vision," Proceedings of the Royal Society B: Biological Sciences, vol. 278, pp. 2241-2248, August 7, 2011 2011.
[71] T. Stich, et al., "Perception-motivated interpolation of image sequences," ACM Trans. Appl. Percept., vol. 8, pp. 1-25, 2011.
[72] T. Stich, et al., "View and Time Interpolation in Image Space," Computer Graphics Forum, vol. 27, pp. 1781-1787, 2008.
[73] J. Hays and A. A. Efros, "Scene completion using millions of photographs," Communications of the ACM, vol. 51, pp. 87-94, 2008.
[74] S. Baker, et al., "A Database and Evaluation Methodology for Optical Flow," International Journal of Computer Vision, vol. 92, pp. 1-31, 2011.