Review on 2D to 3D Image and Video Conversion Methods

Shweta Patil
Dept of Electronics and Tele-communication,
D.Y.Patil College of Engineering, Akurdi,
Pune, India

Priya Charles
Dept of Electronics and Tele-communication,
D.Y.Patil College of Engineering, Akurdi,
Pune, India

ABSTRACT
We present a survey paper on state of the art methods of 2D to 3D image and/or video conversion. In this modern era 3D hardware popularity is increased but, 3D contents are still dominated by its 2D counterpart. Until now many researchers have proposed different methods to close this gap. Mainly, these conversion methods are categorized in an automatic method and semi-automatic method. In an automatic method human intervention is not involved, where as in semi-automatic method human operator is involved. There are distinct attributes that can be considered during conversion, like for video conversion motion is mostly considered parameter; while for image conversion local attributes of images were considered. Computational time and design cost are the main design metrics that should be considered while designing algorithm.

Keywords
Image conversion, video conversion, 2D to 3D, automatic, semi-automatic.

1. INTRODUCTION
In image processing, 2D image is having only two dimensions height and width, 2D images doesn't have depth. While in a 3D image along with the height and width it is having depth information hence known as 3 dimensional images. Today there is an enhancement in 3D capable hardware such as TVs, Blu-Ray players, gaming consoles, and smart phones and many more. These 3D media gives feeling of immersion or more lifelike viewer experience. But the availability of 3D content is not matching with its production rate. There are two methods for generating 3D contents. First, capture the content directly with multiple cameras and other is take 2D conventional footage and converts it to 3D. Former method gives best results but it can be difficult and expensive as it requires specialized equipment and strong production system. The latter method is difficult but may be cost effective.

A typical 2D-to-3D conversion process consists of two steps: depth estimation for a given 2D image and then depth based rendering of a query image in order to form a stereopair images. The latter step of rendering is well realized and algorithms exist that produce good quality results. While the first step of depth estimation from a single image or video frame is a bit challenging. Depth map is a monochromatic image, where a low intensity indicates a far distance from the camera, while a high intensity indicates a closer distance. Basically there are two approaches of 2D-to-3D conversion; one is automatic and another is semi-automatic. In semi-automatic method a skilled operator assigns depth to various parts of an image or video. Based on this sparse depth assignment, algorithm estimates dense depth over the entire image or video sequence. In an automatic method human intervention is not required, a computer algorithm automatically do whole estimation of the depth from a single image or video. Semi-automatic method is much successful but it is time consuming and costly [2].

The problem of depth estimation from a single 2D image, which is the main step in 2D-to-3D conversion, can be formulated in various ways. As image or video is having different attributes, those have considered by different authors and everyone has developed their own algorithm for conversion. Including depth from defocus [9] [10], depth from perspective geometry [11] [12], depth from models [13], depth from visual saliency [14], depth from motion [15] and so on. In the defocus based approaches to extract the blur information from a single image by measuring the amount of blur and then remaps the blur measures to depth map. In [10], the number of high value wavelet transform coefficients is taken as a measure of blur. Perspective geometry refers to the property that parallel lines in real world tend to converge at a point (vanishing point) in the picture. Generally, the vanishing point has the farthest distance, so we can derive a suitable assignment of depth based on the position of the lines and the vanishing points [11]. The approach of depth models constructs several basic depth models of typical scenes and then blends them together to estimate the depth of real scenarios. Visual saliency is another kind of important depth cues based on the analysis of visual attention and a saliency map acts directly as a depth map [16]. Depth from motion is based on the law that near objects move faster across the retina than far objects for a moving observer, so relative motion provides an important depth cue. However, when the objects are also moving, the law does not apply in many cases, which constraints the utilization of this depth cue in 2D to 3D conversion. In many papers described here machine-learning techniques have been used to automatically estimate the depth map from a single monocular image [1-3].

This paper is organized as follows. In section 2, 2D-to-3D image and video conversion methods are discussed. In section 3, 2D-to-3D image conversion method is discussed. In section 4, 2D-to-3D video conversion methods are discussed. In section 5 all discussed methods are compared. In last section conclusion is made.

2. 2D-TO-3D IMAGE AND VIDEO CONVERSION METHODS
The methods described under this section are proposed both for image and video conversion.

2.1 2D-to-3D Image Conversion by Learning Depth from Examples
Janusz Konrad et al. in [1] presented different approach of learning the 3D scene structure. The proposed method is automatic conversion for images. In this method, they have proposed a simplified algorithm that learns the scene depth from a large database which is having image and depth pairs.

Their proposed method is based on observation that among millions of image + depth pairs available on-line, there likely exist many pairs whose 3D content matches that of a 2D input. Also they have made two assumptions that two images that are photometrically similar are likely to have similar 3D structure i.e. depth [1].Since photometric properties are often correlated with 3D content as depth, disparity. For example, edges in a depth map almost always coincide with photometric edges. They have used machine learning technique in their method. Fig.1 shows block diagram of proposed algorithm. From the database containing image and depth pairs using k nearest-neighbor (kNN) search algorithm *k* image + depth pairs that are matched with 2D query left image are searched. For selecting a useful subset of depth relevant images from a large dictionary is to select only the *k* images that are closest to the input. For this they have used distance function the Euclidean norm of the difference between histograms of oriented gradients. Next using median filtering depth fusion of *k* images is generated.

So this method is a simplified and computational efficient data-driven 2D-to-3D conversion method and has insured its performance against state-of-the-art Make3D algorithm. The proposed algorithm compares in terms of both estimated depth quality and computational complexity. This is valid for indoor and outdoor database. The generated anaglyph images produce a comfortable 3D perception but are not completely void of distortions. With the continuously increasing amount of 3D data on-line and with the rapidly growing computing power in the cloud, the proposed algorithm seems a promising alternative to operator assisted 2D-to-3D conversion.

2.2 Learning-Based 2D-to-3D Image and Video Conversion

Janusz Konrad et al. which presented a previous algorithm [1] again proposed an automatic 2D-to-3D image and video conversion method in [2]. Here they have mentioned two methods: one 2D-to-3D conversion by learning a local point transformation. Second is 2D-to-3D conversion based on global nearest-neighbor depth learning. The latter one is the same that is discussed previously, so here we will discuss former one. As image or video frame is having its attributes at a pixel level that is learned by a point transformation. As soon as the point transformation is learned, it is applied to a monocular image. Thus depth is assigned to a pixel based on its attributes.

A key element here is a point transformation which is used for computation of depth from image or video frame attributes. For estimating transformation training on a ground truth database approach is used.

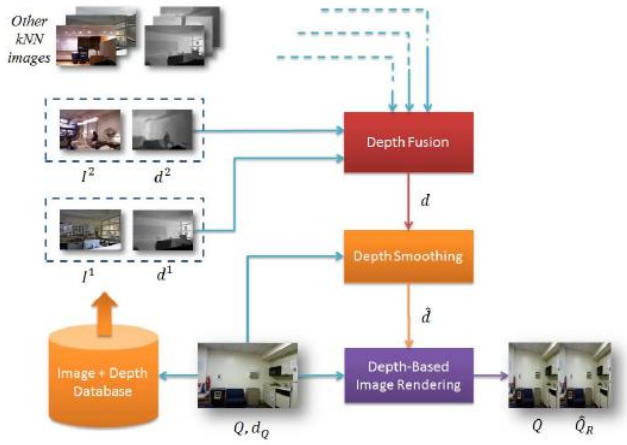

Figure 1: Block diagram of proposed algorithm in [1]

Here point transformation is estimated by training on ground truth. Examples of low-level video attributes that are helpful to compute relative depth of a pixel include color, spatial location, and local motion. Consider there is a dataset having *K* images and depth pairs, from that *I* are the trained images. They have given two examples of such datasets one is Make3D dataset and another is NYU Kinect dataset.

General regression function of learning [2] that maps local features such as color, location, motion is

$$f(\text{color, location, motion}) \rightarrow \text{depth.}$$

The more illustrated form of transformation is:

$$f[color, x, motion] = w_c f_c[color]\, w_l f_l[x]\, w_m f_m[motion] \quad (1)$$

where, f_c is color depth transformation, f_l is location depth transformation , f_m is motion depth transformation, w_c is color depth weight, w_l is location depth weight, w_m is motion depth weight. Here the individual transformations are learned [2].

Motion transformation may contain noise hence bilateral filter is used to overcome noise. Finally all transformations are linearly combined to estimate final depth. Fig. 2 shows a sample video frame with depth maps estimated from color, location and motion cues separately, and the final combined depth map. The point transformation can be learned off-line and applied basically in real time, is the feature of this system. The limitation of this system is, it cannot apply same transformation to images with potentially different global 3D scene structure. To overcome this limitation they have proposed method of global depth estimation in [1].

2.3 2D-to-3D Conversion Algorithm Using Multi-depth Cues

In [3] an automatic algorithm for 2D to 3D conversion is based on multiple depth cues. Here they have considered three distinct depth generation procedures and accordingly 2D scene features one depth generation procedure is executed. The three depth cues they have considered are perspective geometry, defocus, visual saliency and adaptive depth models. Fig 3 shows proposed algorithm for conversion. Before starting this algorithm flow as shown in fig 3, first color image is converted into grayscale image. This grayscale image is then given to the flow.

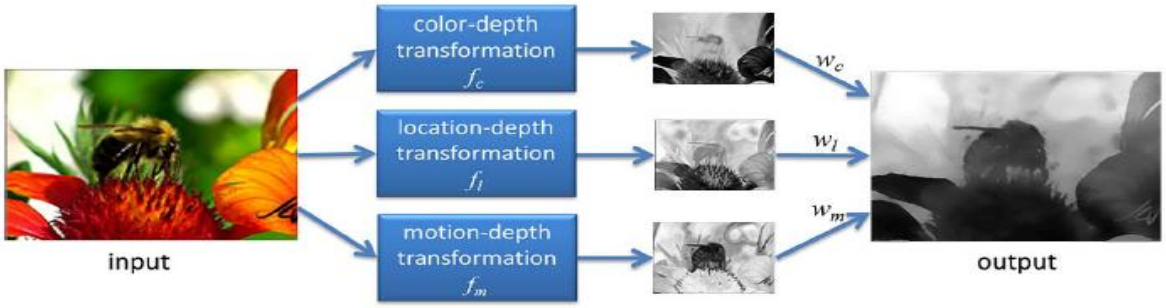

Figure 2: diagram of point transformation in [2]

The very first step is vanishing point detection. For that Canny edge detection is performed on the grayscale image. Then, using Hough transformation the lines in the image and intersections are calculated between the lines detected. If scenes are of perspective geometry, the intersections tend to aggregate to one cluster. A predominant cluster of intersections among all the aggregated clusters are exist, then the scene contains a vanishing point. A cone depth model with the vanishing point is constructed to estimate the depth of the scene. If the scene is not consisting vanishing point then the next is of depth estimation using defocus. Here depth extraction method is based on two dimensional discrete cosine transform (2DCT). For that the input image is partitioned into 8*8 blocks. Then 2DCT in each block is performed. So thus the number of high frequency coefficients which are larger than 1 are calculated. At the end it is remapped to depth range 0-255. The output of the 2DCT is having blocking artifacts, which are overcome by using joint bilateral filter as shown in fig 3. If the image is neither consisting vanishing point nor defocus then the next depth extraction is based on depth models [13]. The three depth models used in [13] are: a spherical surface model, a cylindrical surface and a spherical surface model and a plane and a cylindrical surface model.

2.4 Depth Extraction from Video Using Non-parametric Sampling

Kevin Karsch et al. in [4] presented method to automatically convert a monoscopic video into stereo for 3D visualization. They are taking plausible automatically extracted depth maps at every frame. They have presented a framework for using temporal information for improved and time-coherent depth when multiple frames are available. Also they have collected their own ground truth stereo RGBD (RGB + Depth) video dataset. Working of their system is stared from candidate matching and warping step. From given a database for an input image, high-level image features are computed for each image or frame of video in the database. Then the top K matching frames from the database are selected. One thing here must be noticed that each video in the database contributes no more than one matching frame. By warping procedure the candidates are matched to the structure of the input image. Then using a global optimization procedure the warped candidate images are merged. After that pixel-to-pixel correspondence through Scale Invariant Feature Transformation (SIFT) flow is achieved, which matches per-pixel SIFT features to estimate dense scene alignment. With temporal information e.g., extracted from a video, this algorithm can achieve more accurate, temporally coherent depth.

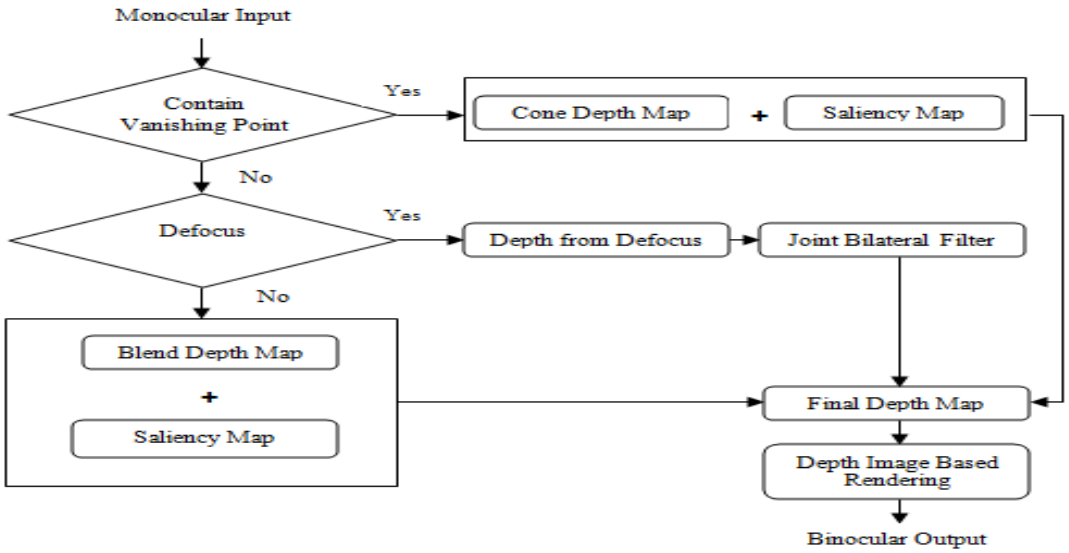

Figure 3: Flow chart of proposed method in [3]

SIFT features are calculated and matched in one-to-many correspond which defines warping function Ψ. For model temporal coherence, first by computing per-pixel optical flow for each pair of consecutive frames in the video can be used to automatically generate the depth maps necessary to produce the stereoscopic video. To avoid generating holes at disocclusions in the view synthesis step, here Wang et al.'s technique [17] is adapted and extend. They developed a method that takes as

input a single image and per-pixel disparity values, and intelligently warps the input image based on the disparity that highly salient regions remain unmodified. Their method was applied only to single images; but here it is extended to handle video sequences as well. So the advantages of this system are no requirement of motion parallax or sequence length and more robust. But in case of video more data is required for more comparisons in the candidate search. This method is applicable to arbitrary videos, moving videos, works in cases where conventional depth recovery methods fail.

3. 2D-TO-3D IMAGE CONVERSION METHOD

The method described under this section is proposed for only image conversion. They cannot be applied for video conversion.

3.1 Image Conversion Using Scale-space Random Walks and a Graph Cuts Based Depth Prior

Phan et al. in [5] presented semi-automatic user-defined strokes corresponding to a rough estimate of the depth values in the scene are defined for the image of interest. This proposed system determines the depth values for the rest of the image, producing a depth map that can be used to create stereoscopic 3D image pairs. This method works in a two stage process using the smoothing properties of Random Walks, and the hard segmentation returned by Graph Cuts. Random Walks is the solution to a linear system and has problems preserving strong edges, but Graph Cuts does this well. However, the hard segmentation with Graph Cuts does not respect smooth gradients or fine detail. By combining the two, they have retained strong object boundaries while also allowing for smooth gradients. So the steps followed are; initially depth map using Graph Cuts is generated first with user-defined depth strokes, in order to generate a depth prior. The depths prior, and the same depth strokes, are integrated into Random Walks as an additional feature when determining the edge weights. The merits of Random Walks are combined with Graph Cuts, in order to produce good quality depth map.

Graph Cuts is based on solving the Maximum-A-Posteriori Markov Random Field labeling problem with hard constraints. The solution to the MAP-MRF is finding the most likely labeling for any given pixel from the provided hard constraints. Depth map generation is a kind of multi-label classification problem, but used formulation only provides a binary segmentation. Therefore, each unique user-defined depth value was assigned an integer label. A binary segmentation was performed for each label separately and the maximum flow values for the graph were recorded. Once the segmentation was performed, a depth prior is generated, and that is used to augment the depth map generated in the Random Walks stage. Then for the same user defined strokes now Random walks method was used. By applying Random Walks directly to the image the final depth map was obtained. They have used advanced version of Random Walks known as Scale-space Random Walks. Using this they tried to preserve global image structure. By incorporating two stages final depth was generated with depth prior.

4. 2D-TO-3D VIDEO CONVERSION METHODS

The methods described under this section are proposed for only video conversion. They cannot be applied for image conversion.

4.1 Semi-automatic Stereo Extraction from Video Footage

Guttmann et al. in [6] proposed a semi-automatic system that converts conventional video shots to stereoscopic video pairs is proposed. They have used classifiers that relate local appearance to disparity estimation within a global optimization scheme. This system works as following stages: (1) User scribbles are marked on some of the frames to indicate desired disparity values. (2) The marked disparities are propagated on the frames on which they are drawn. (3) A classifier is trained for every disparity value marked by the user. (4) The classifier is applied to the entire shot and high confidence predictions are recorded. (5) The disparity map of the entire shot is recovered in an optimization process which is constrained by the original scribbles and the high confidence predictions.

Here in this system they have chosen to estimate disparities directly without estimating depth first. Perceived depth is connected to disparity only through limited amount of parameters. Working directly with disparities has some advantages. Depth on the other hand, can range anywhere between zero and infinity and is inversely proportional the disparity. For disparity propagating through optimization they have considered four types equations: (1) Applied soft constraints that encourage the disparity at each location to be similar to the disparity of its spatial neighbors. (2) Encourage continuity over time with relation to the change in disparity expected from the local motion field. (3) Encourage the system to adhere to the scribbles. (4) Encourage the system to respect the results of the classifiers on anchor points where the confidence values of the classifier are high. For classification they employed a Support Vectors Machine classifier that is trained on the frames marked by the user and then applied to the entire video. This method is having advantages like efficient in terms of human and computer time and therefore cost-effective and accessible. They have applied their method on a variety of broadcast videos like sports videos with long shots, extreme scene and camera motion, on animation sequences, on documentation and on feature films.

4.2 Combining Motion analysis with User Interaction

Miao Liao et al. in [7] presented a semi-automatic system that converts conventional videos into stereoscopic videos by

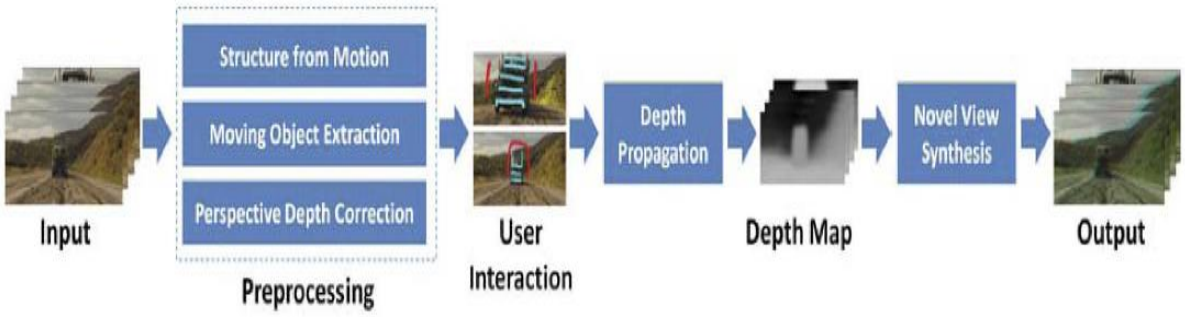

Figure 4: Flow of Proposed Method in [7]

combining motion analysis with user interaction, aiming to transfer as much as possible labeling works from the user to the computer. They have proposed two novel techniques that automatically estimate the 3D information from video sequences. Unlike Structure from Motion (SfM) that requires non-axis camera movement, they have inferred depth information under arbitrary camera/object movement, such as camera pan or zoom, which are frequently used. Here they have given user-friendly interface that requires users to label depth relationship other than depth value on the images. Their User Interface (UI) design benefits from the already defined 3D cues by the preprocessing of movement, providing the users with a more intuitive and less labor intensive UI environment. In the worst case that none of the 3D cues can be inferred in the preprocessing step, their labeling can still simulate the direct depth labeling under the same amount of manual work. They have formulated the sparse to dense depth propagation as a quadratic programming problem that could elegantly integrate both relative and absolute depth constraints. Fig.4 shows flow of the system. In the preprocessing step, the input image sequence is first passed through three individual automatic modules: structure from motion, moving object segmentation (MOS), and perspective depth correction (PDC). The SfM algorithm is applied to the input image sequence with dominant rigidly moving objects to recover a sparse set of 3D points. The MOS module is used to automatically segment the foreground, it is particularly effective in a follow shot in which the foreground is relatively static and the background is rapidly changing. Finally, the PDC module inspects the size change of an object's image to estimate relative depth changes between frames. After automatic processing, the users are presented with images showing area with known depth. If there are still undefined regions, the users need to label them in some key frames by simple scribbling. The user's input as well as all the automatically calculated depth cues will be integrated in a quadratic programming framework to generate dense depth maps for all frames. So this is a hybrid framework to semi-automatically convert conventional videos to stereoscopic videos.

4.3 Video Conversion Based on Depth Propagation from Key-frames

Guo-Shiang Lin et al. in [8] proposed a key-frame selection algorithm which is divided into steps of shot change detection and key-frame assignment. As this is semi-automatic it is up to user to identify and assign frames whose local object motions the weighting parameter, determined by the distance of the current-frame with respect to the front key-frame. The closer the distance is the largerα is. For better improvement this fused depth maps are filter through tri-lateral filters.

are too large to make depth propagation successful as the new key-frames. Hence here color propagation procedure is performed. Then calculate its mean square error with respect to the original image. When the mean square error MSE_i for the i^{th} frame is greater than a pre-determined threshold T_1, it is then identified as a new key-frame [8],

$$keyframe\ assignment = \begin{cases} true & MSE_i \geq T_3 \\ false & Otherwise \end{cases}$$

Fig.5 shows overall flow of proposed system. Here two-pass depth propagation used for depth calculation which is shown in fig 6.

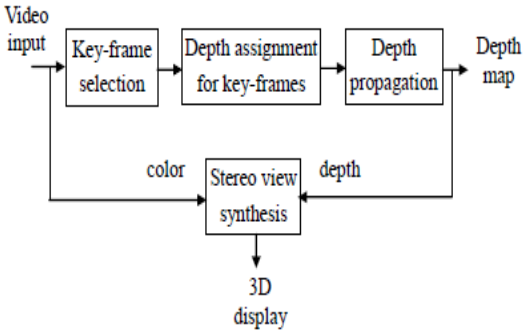

Figure 5: System flow diagram of [8]

The very first step is of initial bi-directional depth propagation. The first pass processing uses the algorithm in [10] to get initial depth maps for each intermediate frame, denoted as $ID_t^{fw/bw}$, where t is a time index and *fw/bw* represent forward and backward respectively. The second pass is erroneous depth block correction. In this pass, erroneous depth blocks in which, $ID_t^{fw/bw}$ are identified by examining their corresponding color min Sum of Absolute Difference (SAD) values in motion vectors determination against a threshold. If the color min *SAD* exceeds, then the corresponding depth block is identified and subject to refinement. After two-pass depth propagation, the depth maps still contain some noise and holes. Therefore in post processing applying techniques like component labeling, region removal, hole filling, etc., to refine the foreground shapes. Finally, using linear weighting scheme bi-directionally propagated depths are fused. In linear weighting scheme α is

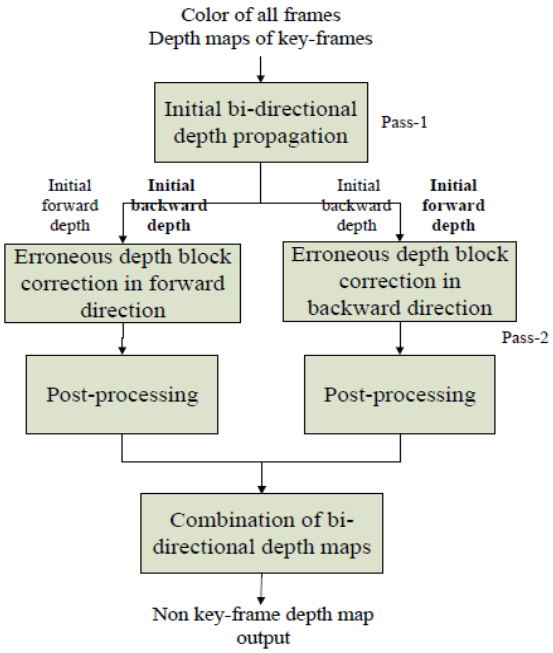

Figure 6: Depth Estimation Flow in [8]

5. COMPARISION OF METHODS

Table I compares discussed various methods on the basis of type of input, whether input to the system is only image or video frame or image and video both is applicable. used algorithm, conclusion and future scope. Table shows that using distinct attributes of image and video frame, 2D to 3D image and video conversion has been done.

6. CONCLUSION

In this literature review paper, we discussed the most recent conversion methods of 2D to 3D image and/or video. There is not fixed steps for doing conversion but taking image or video frames distinct features it is done. Hence 2D to 3D conversion broadly subdivided like depth using motion, depth using visual saliency, depth using perspective geometry etc. Depth using motion is mostly suitable for video conversion. The methods using perspective geometry are more suitable for image conversion. We have covered most of them here. As there is urgent need of 3D contents, using this methods it is possible to close gap between 2D and 3D. Though these methods give good results, still there is room for improvement in future. The discussed methods here are computer vision algorithms, so hardware implementation of them to reduce design metrics and for real time application can be considered as future scope.

Table 1: Summary of Discussed Methods

Type of Input	Algorithm	Conclusion	Future Scope	References
Image and Video	Knn search algorithm	Global method performs better than state of art methods in cumulative performance	Can be work on quality of output image	[1]
Image and Video	local attributes of image or video frames	Local method is outperformed by other algorithms, fast	Can be apply for global images	[2]
Image and Video	Perspective geometry, defocus, visual saliency and depth models	Detect depth cues from image and apply relevant depth extraction method. It gives good quality results	Real time application	[3]
Image and Video	Motion estimation and optical flow	Qualitatively good results, suitable for structure from motion, motion parallax	Ruduce design metrics	[4]
Image	Random walk and Graph cut segmentation	Perceived depth for objects can be directly controlled	Extend for video	[5]
Video	Local saliency and local motion	This method works good for sport videos, extreme scene, camera motion, animation sequences	Extend to automatic	[6]
Video	Structure from motion	Low quality depth map can generate satisfactory results, proposed a new user interface	Extend to automatic	[7]
Video	Motion Estimation	Better 3D perception result, flexible, able to solve depth artifacts caused by background disocclusion	Better motion estimation algorithm	[8]

7. REFERENCE

[1] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in Proc. IEEE Comput. Soc. CVPRW, Jun. 2012, pp. 16-22.

[2] J. Konrad, M. Wang, and P. Ishwar,C.Wu,D. Mukharjee, "Learning based, automatic 2D-to-3D image and video conversion," in Image Processing IEEE Trans on,vol.22,no.9,pp.3485-96,Sept. 2013 Jun. 2012, pp. 16-22.

[3] P.Ji,L.Wang,D. Li, M. Zhang, "An automatic 2D to 3D conversion algorithm using multi-depth cues," IEEE Conf.Audio,Language and Image Processing,pp.546-50,July 2012.

[4] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in Proc. Eur. Conf. Comput. Vis., 2012, pp.775–788.

[5] R. Phan, R. Rzeszutek, and D. Androutsos, "Semi-automatic 2D to 3D image conversion using scale-space random walks and a graph cuts based depth prior," in Proc.18th IEEE Int. Conf. Image Process., Sep. 2011, pp. 865– 868.

[6] M. Guttmann, L. Wolf, and D. Cohen-Or, "Semi-automatic stereo extraction from video footage," in Proc. IEEE Int. Conf. Comput. Vis., Oct.2009, pp.136–142.

[7] M. Lio, J. Gao, R. Yang, and M. Gong,"Video stereolization: Combining motion analysis with user interaction," IEEE Trans.Visualizat. Comput. Graph.,vol. 18,no. 7,pp. 1079-1088, Jun.2012.

[8] G.Lin,J. Huang,W. Lie. "Semi-automatic 2D-to-3D video conversion based on depth propagation from key-frames," Image PracessingIEEE International Conf on,pp.2202-06, Sept 2013.

[9] J. Ens and P. Lawrence, "An investigation of methods of determining depth from focus," IEEE Trans. Pattern Anal.Mach. Intell., vol. 15, no. 2, pp. 523–531, 1993.

[10] S. A. Valencia and R. M. Rodriguez-Dagnino,"Synthesizing stereo 3D views from focus cues inmonoscopic 2D images," in Proc. SPIE, 2003, vol. 5006, pp.377–388.

[11] S. Battiato, S. Curti, M. La Cascia, M. Tortora, and E.Scordato, "Depth map generation by image classification," inProc. SPIE, Apr. 2004, vol. 5302, pp. 95–104.

[12] X. Huang, L.Wang, J. Huang, D. Li, and M. Zhang, "Adepth extraction method based on motion and geometry for2D to 3D conversion", in 3rd Int. Symp. Intell. Inf. Technol.Appl., 2009, pp. 294–298.

[13] K. Yamada and Y. Suzuki, "Real-time 2D-to-3Dconversion at full HD1080P resolution", the 13th IEEEInternational Symposium on Consumer Electronics, 2009,pp.103-107.

[14] C. Huang, Q. Liu and S. Yu, "Regions of interestextraction from color image based on visual saliency",Springer Science Business Media, 2010.

[15] E. Imre, S. Knorr, A. A. Alatan, and T. Sikora,"Prioritized sequential 3D reconstruction in video sequences of dynamic scenes," in IEEE Int.Conf. Image Process. (ICIP),Atlanta, GA, 2006.

[16] J. Kim, A. Baik, Y. J. Jung, and D. Park, "2D-to-3Dconversion by using visual attention analysis," in Proc. SPIE7524, Feb. 2010, 752412.

[17] O. Wang, M. Lang, M. Frei, A. Hornung, A. Smolic, andM. Gross, "StereoBrush: Interactive 2D to 3D conversion using discontinuous warps," in SBIM, 2011.