AUTOMATIC 2D-TO-3D CONVERSION OF SINGLE LOW DEPTH-OF-FIELD IMAGES

Serendra Reddy

Supervisor: Associate Professor Fred Nicolls

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical Engineering

UNIVERSITY OF CAPE TOWN

2016

Abstract *This research presents a novel approach to the automatic rendering of 3D stereoscopic disparity image pairs from single 2D low depth-of-field (LDOF) images. Initially a depth map is produced through the assignment of depth to every delineated object and region in the image. Subsequently the left and right disparity images are produced through depth image-based rendering (DIBR). The objects and regions in the image are initially assigned to one of six proposed groups or labels. Labelling is performed in two stages. The first involves the delineation of the dominant object-of-interest (OOI). The second involves the global object and region grouping of the non-OOI regions. The matting of the OOI is also performed in two stages. Initially the in focus foreground or region-of-interest (ROI) is separated from the out of focus background. This is achieved through the correlation of edge, gradient and higher-order statistics (HOS) saliencies. Refinement of the ROI is performed using k-means segmentation and CIEDE2000 colour-difference matching. Subsequently the OOI is extracted from within the ROI through analysis of the dominant gradients and edge saliencies together with k-means segmentation. Depth is assigned to each of the six labels by correlating Gestalt-based principles with vanishing point estimation, gradient plane approximation and depth from defocus (DfD). To minimise some of the dis-occlusions that are generated through the 3D warping sub-process within the DIBR process the depth map is pre-smoothed using an asymmetric bilateral filter. Hole-filling of the remaining dis-occlusions is performed through nearest-neighbour horizontal interpolation, which incorporates depth as well as direction of warp. To minimising the effects of the lateral striations, specific directional Gaussian and circular averaging smoothing is applied independently to each view, with additional average filtering applied to the border transitions. Each stage of the proposed model is benchmarked against data from several significant publications. Novel contributions are made in the sub-speciality fields of ROI estimation, OOI matting, LDOF image classification, Gestalt-based region categorisation, vanishing point detection, relative depth assignment and hole-filling or inpainting. An important contribution is made towards the overall knowledge base of automatic 2D-to-3D conversion techniques, through the collation of existing information, expansion of existing methods and development of newer concepts.*

Acknowledgements

I would like to thank my supervisor Fred Nicolls for his valuable guidance and input into the structure and direction of the thesis.

I would like to thank my family Anupa, Leya and Zaydin for their motivation, support and patience. I would like to thank especially my parents Thayalan, Dolly, Sarojini and Vasi for their unwavering dedication and encouragement over the years.

I would also like to acknowledge the financial assistance of the Durban University of Technology.

CONTENTS

I. INTRODUCTION

This chapter provides a brief overview of the underlying concepts and techniques employed in 2D-to-3D conversion and motivates the need for improved and augmented methodologies in this field. The challenges of the research are highlighted and a summary of the proposed method as well as the novel contributions made in this research are presented. An outline of the thesis concludes this chapter. This research proposes a novel method for the automatic generation of disparity image pairs from single 2D low depth-of-field (LDOF) images.

A. Preamble

One of the popular applications of 2D-to-3D conversion is its use in 3D cinema and 3D television (3DTV). Even with the surge in these arenas over the past decade, virtually the entire corpus of previous and current cinema and television has been and is still being shot using standard monocular 2D cinematography. Furthermore, movies and programmes destined for 3D platforms[1] currently require significant portions to be filmed using monoscopic cameras, owing mainly to the technical complexity and limitations of filming with stereoscopic cameras [2]. As a consequence there is a significant demand for highly accurate and efficient 2D-to-3D conversion techniques to firstly, help in ameliorating the present shortage of high-quality 3D content by converting some of the vast amount of existing content and secondly, assist in overcoming some of the complications associated with stereoscopic filming and the creation of modern 3D cinema.

3D cinema and 3DTV is by no means an absolute necessity for cinema or television. However, for many, it is seen as providing a more compelling, emotional, psychological and realistic experience, owing primarily to having a sense of perceived immersion into a dynamic image environment and the ability to engage and discern tangible solid objects in the three spatial dimensions, all with the awareness of the fourth dimension, time. Nevertheless, it must be emphasised that an artistically or technically bad 2D film or programme may not, in any way, be transformed into a good product by converting it into 3D. On the other hand, a good 2D film or programme may be transformed into a bad quality product if inaccurate or inadequate 3D conversion techniques are employed.

Stereoscopic or 3D cinema had a brief appearance in the early1920s, late 1930s and again in the early 1950s [3] and even though sporadic attempts at resurrection were made in every

[1] A 3D display system or platform, as defined in this research, is one which provides additional information to an observer as a consequence of the observer being endowed with two eyes and stereoscopic vision (in the very small minority, the former does not necessarily imply the latter) [1].

subsequent decade it went largely unnoticed for over half a century. However, this changed around 2005 [4] where, owing mainly to the divergence of technologies and products with consumers and content, a resurgence began and what ensued, year on year, since then is observed to have followed a "Moore's law"-type trend for both the cinematic and home entertainment 3D markets [5].

Two prominent films of recent, *Titanic* and *Jurassic Park*, have been transformed, with a relatively high degree of success, from 2D into 3D [6, 7]. However, these exercises proved to be extremely manually intensive, requiring a colossal army of 3D modellers, rendering artists, stereoscopic specialists and video engineers and all at a phenomenally exorbitant cost.[2]

The fundamental problem in 2D-to-3D conversion is that given only limited information in two dimensions, a third dimension, depth, has to be extrapolated. Simians, including humans, observe or rather experience the world in three dimensions. This visual experience is principally a consequence of stereoscopy or stereopsis i.e. the propensity to perceive and appreciate depth by being able to differentiate relative distances to objects as well as distinguish the position of individual objects in relation to other objects in a scene. This stereopsis, as it relates to primates, is possible owing largely to the disparity between the right and left retinas, whose photoreceptors simultaneously capture two separate 2D images of the same scene from slightly different viewpoints or angles. These images are then processed, merged and interpreted by the brain as a single 3D scene [8, 9].

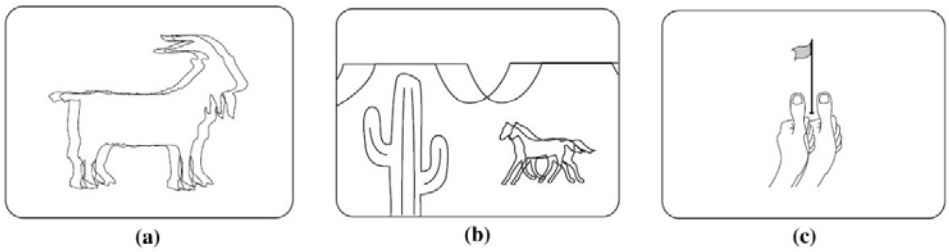


Fig. 1 Stereopsis [10]. (a) Retinal disparity. Overlaid left and right retinal images; (b) Focus on cactus in front. Positive parallax. Horse and mesas will appear blurry; (c) Focus on flag in distance. Negative parallax. Thumbs will appear blurry. In the case of (b) and (c), should focus be maintained and only one eye kept open, the parallax will be negated; however, the defocus will remain.

Euclid, the father of modern geometry and the architect behind the foundations of "stereoscopic art", said nearly 2000 years ago, "To see in relief is to receive by means of each eye

[2] The conversion of Titanic into 3D took 14 months to complete and required a team of around 450 contributors and at a cost of $18 million [6]. The conversion of Jurassic Park into 3D took 10 months to complete and required a team of around 700 contributors and at a cost of $10 million [7].

the simultaneous impression of two dissimilar images of the same object [11].” Our own awareness of spatial relations is based on the innate ability of the human visual system (HVS) to capture, process and understand the Euclidean structure of our 3D environment [12-14]. This is owing, in the main, to a number of reliable and cooperative visual monoscopic and stereoscopic heuristics or depth cues [15].

Even though respected mathematicians like Euclid, Da Vinci and Kepler understood we see different images of our environment through each eye, it was only given credibility in 1838 when Wheatstone described this phenomenon as being stereopsis: the stereoscopic receiving of information, owing to retinal disparity and the subsequent perception of three dimensions [10, 16]. This is illustrated in Fig. 1(a).

In addition to retinal disparity, another physiological cue present in stereoscopy is vergence. The former is the positional difference between the two retinal images of a scene point. This is owing to the distance between the two eyes (average approximation of 64 mm in humans). The latter is the angle between the line-of-sight of both eyes [17].

A crucial component to 3D perception is having access to data acquired simultaneously from both eyes. This allows for the brain to differentiate relative distances to objects as well as distinguish the position of individual objects in relation to other objects within a scene. Although visual information is received simultaneously by each eye, the information is different for each eye, owing to the angular disparity (vergence) between the eyes. The brain extrapolates the third dimension, depth, through triangulation of this discrepancy in the relative 2D positions of the points in the two disparity images; a consequence being depth perception [18].

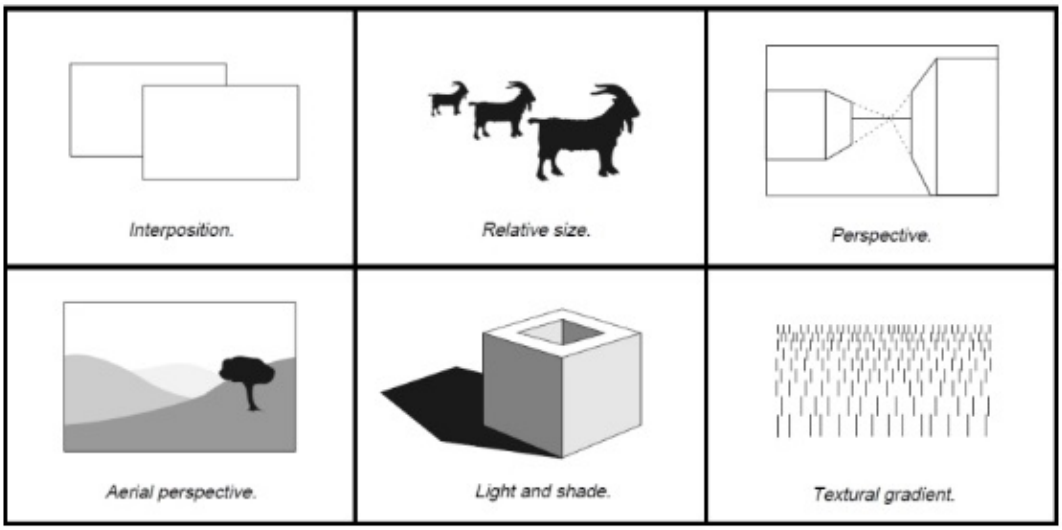


Fig. 2 Monoscopic depth cues [17].

In the absence of stereoscopic or binocular depth cues the brain is nevertheless capable of perceiving depth through the inference of monoscopic or monocular depth cues. There are

nine individual basic monoscopic depth cues [10, 15, 17, 19]; these include (i) motion parallax, (ii) depth-of-field (focus and defocus), (iii) linear perspective, (iv) interposition or occlusion, (v) relative height, (vi) light and shade (shading), (vii) relative or known size, (viii) atmospheric scattering or aerial perspective and (ix) textural gradient. Fig. 2 provides an illustration for six of these cues.

Motion parallax refers to the relative motion of objects across the retina. The depth value of an object is assumed to be proportional to the motion vector, whereby the faster the object moves, the closer it is to the observer and vice-versa, for slower moving objects [20].

Focus and defocus cues refer to the ability to infer depth based on the focus-defocus phenomena whereby objects will exhibit a certain amount of blur (out of focus) relative to their position very near to, or very far from, the object in focus – this is known as depth-of-field (DOF)[3].

Linear perspective is the appearance of depth on a 2D plane through the convergence of imaginary straight lines, whereby the more the number of lines converge, the more it will have the appearance of being further away [21-23]. The lines may be inferred from objects and artefacts such as buildings, walls, roads, street markings and railway tracks.

Occlusion or interposition refers to the simple observation that when an object is fully or partially hidden the occluded (hidden) object is considered to be further away than the occluding object that is blocking or covering it.

Relative height refers to the distance from the top or the bottom of the border of the actual image and not as a size dimension. It may be observed that objects towards the top of an image are deemed to be further away than objects closer to the bottom of the image. This may be associated with Gestalt principles [24].

Light and shade cues are commonly called shape-from-shading (SFS) and refers to the inferring of 3D structure based on the brightness and gradual variation of the shading of objects in a scene [25-28].

Relative or known size is based again on observation whereby objects appear smaller when they are farther away and larger when they are closer. Knowledge is necessary to make a judgment about the distance of familiar objects. For example, a car seen at a great distance is interpreted to be far away and not necessarily small.

Atmospheric scattering (haze) or aerial perspective refers to the observation that scattering of red light by the turbid medium in the atmosphere produces a partially polarised bluish haze or tint that results in objects in the distant appearing less contrasted than nearer objects [29]. The

[3] In the HVS the accommodation of the eye is to focus on a given OOI and the brain subsequently stimulates the 3D emersion effect through the inference of the relative depths of the surrounding regions depending on the degrees of defocus. In the case of a LDOF image captured by a camera the effect of the focus and defocus is associated with the setup of the optical system of the capture device.

same holds true for the observation of objects in fog, mist and smoke. Although each medium differs in size, shape, material and concentration of atmospheric particles the effect is collectively referred to as haze.

Textural gradient refers to the observation that the texture of the surfaces of certain objects appear distorted in a way that actually reflects both the direction as well as the inclination of the surface [30, 31].

These monoscopic depth cues are the reason why it is possible, with reasonable accuracy, to gauge depth using just one eye. Therefore, it may be possible to estimate the depth of objects from the camera, as well as the distance relationship every object has relative to each other, with a combination of these monoscopic artefacts.

3D perception and emersion, as experienced by the "mind's eye", is a consequence of the simultaneous coalescing and interpretation by the HVS of all these monoscopic and stereoscopic artefacts and phenomena. It is nevertheless possible, with reasonable accuracy, to gauge depth using just one eye i.e. absent of stereoscopic disparity. A monocular 2D image or video sequence is in principle equivalent to viewing a scene through only one eye. Based on this understanding it may therefore be possible for a for a machine to extract a reasonable amount of depth, shape and relative distance information from 2D images and sequential video through the mathematical translation of these monoscopic artefacts. All of these monocular depth cues may in some way contribute towards estimating a scene's 3D structure. However, owing to the variety of scenes in random images, all of these depth cues are individually imperfect and in some scenarios may be non-existent or provide an inadequate amount of information to allow for depth computation.

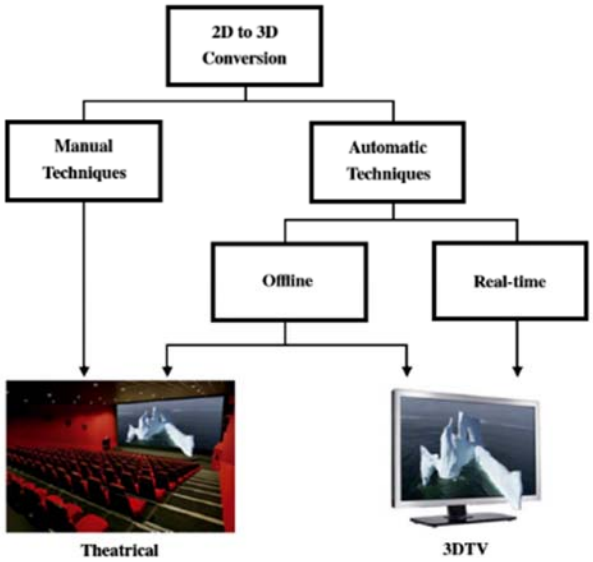
Moreover, when considering the automatic conversion of single 2D images into 3D it is understood that no motion or a priori information exists. As a consequence, some of the abovementioned depth cues become more difficult to extrapolate than others. Relative size, for example, may only be applied based on the identification or description of objects, together with information on the normal sizes of those respective objects. Also, occlusion may only really provide information on which objects are in front of other objects. However, absent of motion, no information pertaining the actual depth of those objects from the camera or the relative distances between each object may be extracted. Textural gradient is also limited to only specific kinds of images and as with relative size, requires a priori knowledge to be applicable.

Although relative depth may be gauged when viewing monocular images and video, the actual experience of depth in these scenarios is attenuated. This is owing to the absence of information one acquires from the matching left or right disparity image, which is present in stereoscopy. It is nevertheless possible to create this perception of depth by simultaneously presenting a matching disparity image to the HVS, as is experienced for example in 3D cinema or TV.

There are currently four types of 3D display formats. These include anaglyph, polarised, active shutter and autostereoscopic. Irrespective of which display format is used the images have

to nevertheless be separated by either specialised glasses or screens into their individual disparity images and then independently but simultaneously presented to each eye. Based on this understanding, 2D-to-3D conversion may then be described as the extrapolation of the third dimension, depth, from a single 2D image or sequence of monocular video frames and the subsequent rendering of a disparity (separate laterally shifted left and right eye) image pair.

There are three approaches to 2D-to-3D conversion [32]. These include manual, semi-automatic (human-assisted) and automatic (unsupervised). A comparison between each of these methods together with their trade-offs are provided in Fig. 3.



Conversion	Manual	Automatic	
		Offline	Real-time
Speed	-	+/-	+
Quality	+	+/-	-

Fig. 3 Overview of the 2D-to-3D conversion approaches with comparison of trade-offs [32].

The manual conversion techniques are mostly utilised for cinematic productions. This is owing, in the main, to quality concerns and significant attached costs. The semi-automatic model is whereby a part of the 3D conversion process is done automatically, followed by "manual" corrections done by hand. Even though this scheme may significantly reduce the overall time consumption, when compared to the fully manual conversion scheme, a significant amount of human interaction is nevertheless necessary for the completion of the 3D conversion. To economically convert the vast amounts of available 2D material into 3D a fully-automatic conversion system is desired [19].

There two approaches to automatic or unsupervised 2D-to-3D conversion viz. real-time and non-real-time (off-line). The real-time automatic techniques are largely employed in the 3DTV market which, owing to a surge in consumer interest, has warranted a significant escalation in new research and products. Although the market trend shows an increase in demand for these real-time products there are nevertheless several inherent problems, including cost, poor quality and viewer

discomfort. On the other hand, the off-line automatic 2D-to-3D conversion models are seen as the compromise between speed and quality.

B. Research Challenges

Over the past two decades there have been numerous approaches to autonomous 2D-to-3D image and video conversion [19]. These methods may be broadly divided into two categories. The first category involves the direct synthesis of disparity stereo image pairs based on the exploitation of the motion parallax (MP) by estimating planar transformations in sequential monocular video [33-35].

The second category involves the recovery of depth information from monocular images and sequential video and the subsequent synthesis of the disparity stereo image pairs from this knowledge of depth. These are referred to as depth image-based rendering (DIBR) techniques [36].

The MP methods are only applicable to specific modes of 2D monocular video capture and are incompatible with single 2D images. The DIBR methods, on the other hand, provide a more generic approach to 2D-to-3D conversion and may be applied to single 2D images. An overview of a standard depth-based 2D-to-3D conversion system is illustrated in Fig. 4.

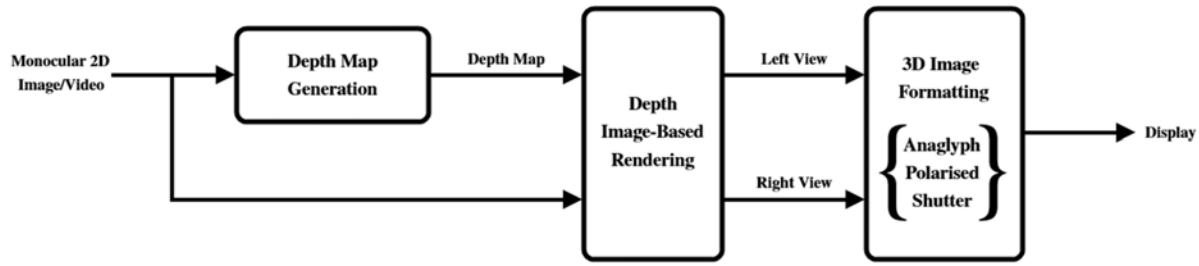


Fig. 4 Overview of the 2D-to-3D Conversion Process [37].

An automatic depth-based 2D-to-3D conversion system must be able to firstly, extract information from an inputted 2D image or sequential video frames absent of any a priori knowledge of objects and relative depths of objects in the scene and secondly, directly compute the depth dimension [38, 39]. Challenges to these types of autonomous 2D-to-3D conversion models exist on several levels.

Fig. 5 An input image and its estimated depth map [40]. In this case the higher the intensity of the pixel (whiter) implies a closer relative depth to the camera.

The association of a depth value to every pixel in an image is termed a depth map and is commonly represented in the form of a grey-level intensity array. This is illustrated in Fig. 5. For the assignment of depth to be effective it is necessary to account for the multiple objects and sub-regions present within a scene; all of which have to be meticulously delineated and categorised. In addition to accurately delineating these objects and regions, objects effectively occluding (in front of and partially blocking) other objects need to be identified. This is essential, since every pixel has to be grouped and assigned an appropriate and relative depth value.

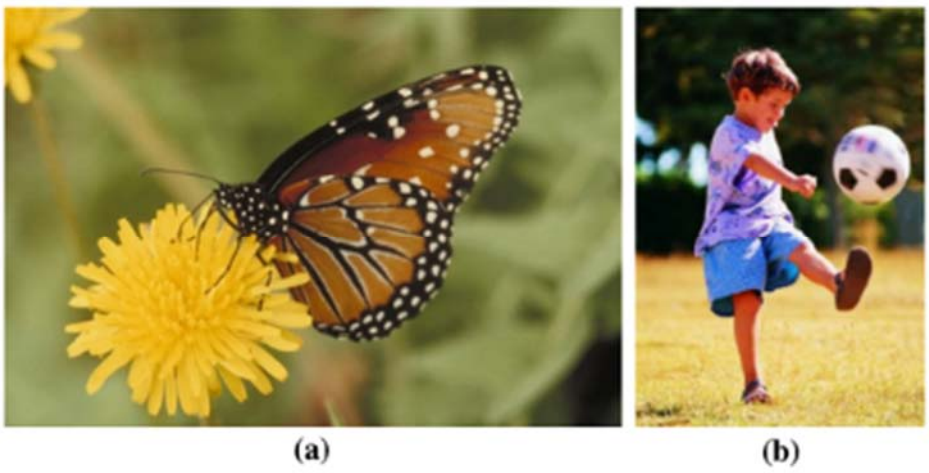


Fig. 6 Region-of-interest [41, 42]. (a) OOI; (b) OOI is a subcomponent of a larger ROI.

The emphasis of this research is on the automatic 2D-to-3D conversion of single 2D low depth-of-field images. Image or video capture is broadly categorised into two depth-of-field (DOF) techniques. These include high DOF and low DOF (LDOF). The term DOF in photography or cinematography refers to amount of focus (sharpness) or defocus (blurriness) present in an image. Two examples are illustrated in Fig. 6. The sharp regions are usually

associated with the foreground (FG) region while the blurred regions are typically associated with the background (BG) regions.

One of the key purposes of LDOF image capture is to highlight specific regions in an image where the viewer is expected to pay more attention to. In addition to FG emphasis, objects in the scene of a LDOF image will exhibit a certain amount of blur relative to their position very near to, or very far from, the object or region in focus. This understanding may allow for the possible extrapolation of depth through relative blur comparison.

The FG is commonly referred as the region-of-interest (ROI). Within a ROI there may also exist an object-of-interest (OOI). There are two types OOI scenarios in LDOF images. The first is where the OOI constitutes the entire ROI and the second is where the OOI is a subcomponent of the ROI. These are illustrated in Fig. 6(a) and (b), respectively.

In the case of Fig. 6(b) it can be seen that the boy, ball and part of the ground form the ROI. However, the OOI should only be the boy and possibly the ball. For simple ROI extraction this delineation is not considered to be important. However, for the assignment of relative depth it becomes significant, since the FG ground region forms part of a larger ground region and the depth of the ground incrementally increases away from the camera towards the tree or horizon line while the depth of the boy and the ball as a collective will be at a relatively constant distance away from the camera.

The recognition and visual delineation of the ROI and OOI play a vital role in both multimedia entertainment as well as machine vision. There are no optimal solutions for the unsupervised extraction of the ROI or matting of the OOI. Motion represents one type of depth cue and may provide a means of reasonably estimating the location of an OOI in an image as well as allow for the possible extrapolation of relative depth based on motion estimation. In the absence of motion, as in the case of single 2D LDOF images, segmentation and relative depth estimation will need to rely solely on other monoscopic depth cues.

In 2D-to-3D conversion the segmentation or categorisation of the regions in an image is referred to as labelling [43]. Some of the most common regions include the OOI, ground region, background and skyline. Emphasis in an image is usually placed on the OOI and the accurate delineation of this region is a crucial component in the efficacy of the final 3D product. The process of extracting or differentiating the OOI from the rest of the image is referred to as image matting [44, 45].

Although an accurate depth map is crucial, it is nevertheless only the first step in a three-step process required for an end-product stereo image. Step two, which is commonly referred to as depth image-based rendering (DIBR) [46-48], is the process whereby one or more disparity images are generated using the original image and its associated depth map. In the case of 3D, DIBR refers to the synthesising of two simulated stereoscopic partner images that represents the view of the same image or scene from slightly different angles [49].

To create this so called “second view” or “second eye” the pixels representing objects closer to the camera, as determined by the depth map of the image, need to be moved either left or right, in front of the background pixels. This pixel location reassignment process is commonly referred to as 3D warping. A consequence of warping is that some pixels will become occluded and other pixels will be revealed, resulting in missing pixels or “gaps” appearing in the background. This is illustrated in Fig. 7.

Fig. 7 DIBR Image Warping. (a) Original image; (b) Left warped; (c) Right warped.

From the several associated complexities within the DIBR process, an immediate and significant concern is on how to deal with the newly exposed "dis-occluded" regions appearing in the rendered images. This sub-step of the DIBR process, which involves the predicting and subsequent filling of these missing pixels, is referred to as hole-filling.

The problem with these dis-occlusions is that even with a highly accurate depth map there is no precise means of determining the values of the revealed pixels, since the depth map only provides depth information and not intensity nor lateral nor angular information of the occluded regions. In most cases, specifically single images and static video frames, these values may only ever be predicted within a certain probability, since they never actually exist.

These newly created holes need to be resolved effectively so as not to significantly distort the quality of the synthesised disparity images. This is a crucial issue, as it has direct bearing on the overall accuracy and experience of the final product and is, therefore, among the most difficult and challenging tasks present in 2D-to-3D conversion [47].

The rendering of a disparity image pair in effect concludes the 2D-to-3D conversion process. However, for the 3D effect to be experienced it is necessary to for these stereo images to be exhibited in a relevant display format. In other words, once the aforementioned matching left and right disparity images are extrapolated then, based on the prescribed display format, both these images need to somehow be presented simultaneously to each eye of the observer. This is usually achieved through means of either filtered glasses or filtered display screens that allow for the disparity image pairs to be simultaneously projected, separated and directed into each retina of the observer.

There are currently four popular modes employed for the presentation of 3D images and video. These include anaglyph, polarised, shutter and autostereoscopic, with anaglyph being the simplest and most economical method [50]. A sample anaglyph image is illustrated in Fig. 8.

Fig. 8 Anaglyph (red-cyan) image of a scene from *The Lord of the Rings* (Courtesy of New Line Cinema).

Anaglyph images are produced by superimposing two individually coloured (typically red and cyan) disparity images into a single image. For the 3D effect, this image may then be viewed through glasses having similar coloured filters. Although anaglyph provides a relatively uncomplicated means of achieving the 3D visual experience, there nevertheless exist several deficiencies, the most prominent being colour distortions, retinal rivalry and ghosting. For a consistent and comfortable 3D visual experience these inherent problems need to be minimised in the final product.

From the above discussion it is evident that several difficulties exist at different stages of a 2D-to-3D conversion process. Moreover, the complexities increase when firstly, the model is required to function automatically (unsupervised) with no a priori knowledge and secondly, when single 2D images are concerned.

C. Summary of the Proposed Approach

A 2D-to-3D conversion system is broadly described as having three processes. This includes depth map generation, depth image-based rendering (DIBR) and stereo image presentation.

A 2D image represents a scene captured from a single viewpoint and the 2D-to-3D conversion of this image therefore involves the rendering of two images representing the respective scene being observed from two slightly different lateral viewpoints. In the proposed model the generation of these left and right disparity images is achieved through the use of DIBR. The stereoscopic image pair is rendered through the warping of the original 2D LDOF source image

using an extrapolated depth map of the image. The depth map is autonomously produced by firstly, accurately delineating and segmenting every object and region in the image and secondly, extrapolating the depth (the third dimension) for every pixel in the image using only the available two-dimensional pixel information.

The generation of the depth map is only the first part of the automatic 2D-to-3D conversion system. The actual 3D conversion occurs when the stereoscopic disparity images are produced through the DIBR process. The initial disparity is created through the warping (lateral shifting) of the objects and regions in the image according to their associated relative depth values in the extrapolated depth map. A notable effect of the warping process is the appearance of several dis-occlusions in both the left and right disparity images. This is owing to some of the shallower depth objects and regions shifting and occluding some of the deeper depth objects and regions and in the process revealing non-existent pixel gaps in the BG. The techniques employed to resolve these dis-occlusions is termed hole-filling or image inpainting.

From this discussion, the efficacy of the final rendered stereoscopic image pair is dependent on both the accuracy of the warping, which is inherently dependent on the accuracy of the depth map, as well as the eradication or minimisation of the discontinuities associated with the warping sub-process.

The following breakdown provides a high-level overview of the proposed automatic 2D-to-3D conversion system for single 2D LDOF images:

Stage 1: ROI extraction;
Stage 2: Low depth-of-field image classification;
Stage 3: OOI matting;
Stage 4: Global segmentation and labelling based on Gestalt principles
Stage 5a: Assignment of relative depths using Gaussian re-blur analysis;
Stage 5b: Assignment of relative depths using a gradient-plane;
Stage 6: Stereoscopic image synthesis using DIBR; and
Stage 7: Anaglyph image presentation.

Stage 1 of the proposed model involves the extraction of the region-of-interest (ROI) from the LDOF image. Although there exist several objects and regions within a LDOF image, at this stage the emphasis is placed on the delineation of the foreground (FG) or the so-called ROI from the background (BG) regions. This is achieved by correlating the edge, gradient and higher-order statistics (HOS) saliencies in the image. The edge locations are estimated by combining both RGB edges with the lightness layer edges of the CIEL*a*b* colour space using the Sobel and Canny methods. This edge combination is shown to be the most robust for LDOF images. The gradients are estimated using the Sobel method and for the HOS analysis, the 4^{th}-order moment is considered. The ROI is produced by initially thresholding the gradient and HOS saliency maps

and subsequently correlating and refining the region using edges saliencies together with *k*-means cluster segmentation and CIEDE2000 colour-difference analysis.

Stage 2 of the proposed model involves the classification of the LDOF image. Every LDOF image contains either two or three distinct regions. The two primary regions are the FG or the ROI and the BG and the secondary region is the object-of-interest (OOI) within the ROI. All LDOF images must contain an ROI but not necessarily an OOI. Moreover, if an OOI exists, then it may either be equivalent to the entire ROI or represent a sub-region of the ROI. For the latter, the ROI then contains two regions. These include the in focus OOI and an in focus ground region. Based on this understanding every LDOF image may then be classified by three regions. These are the in focus OOI, in focus ground region and BG. The ROI is determined in the first stage of the proposed model. A rudimentary estimate of the OOI is produced by considering the dominant gradients contained within the confines of the ROI. Subsequently, the in focus ground region is extrapolated using the ROI and OOI. The delineation of these regions then allows for the tri-region classification of the LDOF image.

Stage 3 of the proposed model involves the matting of the OOI from within the ROI. This stage is dependent on the classification of the LDOF image and only applies to scenarios where a distinct in focus ground region is present. For all other scenarios the OOI is either non-existent or equivalent to the entire ROI. The OOI, which is an element of the ROI, describes the region where an observer will focus, or concentrate, more intently and which it is reasonable to assume will have the most influence on the quality of the final 3D product. As such, the accurate delineation of this closed boundary region and the subsequent assignment of relative depth to this region is expected to have the greatest impact on the efficacy of the depth map. Therefore, in the proposed model more prominence is given to the accurate delineation and subsequent extraction of the OOI compared to any other subprocess within the depth map generation stage. The OOI is isolated and accurately matted from the ROI using the dominant gradients, edge saliencies and *k*-means cluster segmentation.

Stage 4 of the proposed model may be broadly described as the complete delineation and subsequent segmentation of the LDOF image into its constituent objects and regions, as well as the labelling or categorisation of these segmented regions within the image. Six Gestalt-based regions are proposed. These include the ground region, the OOI, objects on top of the ground region, objects on the periphery of the ground region, skyline region and objects within the skyline region. Segmentation is performed using *k*-means and binary quantisation clustering.

Stage 5 of the proposed model is responsible for the generation of the grey-level intensity depth map. This stage is split into three substages. The non-OOI region in a LDOF image may be described as either equidistant-based or gradient-plane-based. The first substage involves the sub-classification of the image into one of the two proposed scenarios. This is achieved through the correlation of clusters derived from the k-means and binary quantisation segmentation of the non-

OOI region. The second substage deals with the equidistant-based situation and the third substage accounts for the gradient-plane-based scenario.

Equidistant-based implies that the non-OOI region in the LDOF image are all located at the same relative distance behind the OOI. In this scenario the relative depths of the non-OOI region are determined using depth from defocus (DfD). This is a method of estimating depth based on the relationship between the focused and defocused regions in a LDOF image. In the proposed model, owing to source being single LDOF images, a Gaussian re-blur analysis DfD technique is chosen. This involves firstly, generating a blurred version of the source LDOF image using a Gaussian kernel and secondly, determining the Gaussian defocus blur by calculating the ratio of the gradients between the original and blurred image. To minimise outliers, a shiftable joint bilateral filter is applied to the Gaussian defocus ratio map. Since the edges of objects have a high probability of being the edges of the depth map, more emphasis is given to the edge locations. As a consequence the Gaussian defocus ratio map is filtered a second time using a domain transform edge-preserving filter.

Gradient-plane-based implies that the non-OOI region in the LDOF image contains a distinct ground region as well as permutations of the other four Gestalt-based regions. In this scenario the assignment of relative depths to the objects (including the OOI) and regions in the LDOF image is based on their respective properties as well as their associated location on a gradient-plane. The gradient-plane is constructed using the ground region together with vanishing point (VP) estimation. VP estimation is based on the analysis of the convergence and intersecting of the dominant straight edges in the image directly in the image-plane parametric space.

For both substages the depth map is finalised through the application of a domain transform edge-preserving filter. This is done to minimises some of the irregularities, such as blocky artefacts, discontinuities and outliers.

Stage 6 of the proposed model involves the synthesis of a stereoscopic image pair using depth image-based rendering (DIBR). DIBR is whereby objects and regions in the image are laterally shifted according to their associated depth values in the depth map in order to produce a new image that will appear as if it were captured from a new viewpoint. A consequence of this shifting or warping is some pixels will become “dis-occluded”, creating gaps in the generated image. By generating two disparity image pairs, contrary to a single left or right disparity image, the width of the dis-occlusions is effectively halved, thereby spreading the discrepancies across the images. The dis-occlusions are further minimised by applying an asymmetrical edge-preserving bilateral filter to the depth map prior to the warping sub-process. To resolve the remaining dis-occlusions, hole-filling is subsequently performed using nearest-neighbour horizontal interpolation of the BG regions. The interpolation also takes into account the depth as well as direction of warp. As a means of minimising the striping artefacts, specific directional Gaussian and circular averaging smoothing is applied independently to each view, with additional average filtering applied to the

border transitions. Both the pre- and post-processing smoothing is shown to effectively alleviate a significant amount of the discontinuities while maintaining the subjective view quality.

The rendering of a left and right stereoscopic image pair from a single 2D LDOF image concludes the 2D-to-3D conversion process. However, as a means of demonstrating the efficacy of the algorithm, an additional stage is incorporated into the proposed model.

Stage 7 involves the generation and presentation of a single 3D anaglyph (red-cyan) image based on the merging of the left and right disparity images. Anaglyph images are prone to certain deficiencies, such as colour distortions, retinal rivalry and ghosting effects. To simultaneously minimise these concerns, the proposed algorithm matches perceptual colour appearance attributes in the uniform CIEL*a*b* colour space of the stereo disparity pairs, rather than only direct matching of the CIEL*a*b* or CIE XYZ values. This approach is shown to be more accurately aligned with the colour perception of the HVS.

D. Novel Contributions

The original contributions made in this thesis include:

1. An extremely robust automatic technique for the extraction of the FG or ROI from a LDOF image through the interrogation and correlation of edge, gradient and higher-order statistics (HOS) saliencies. The results show the proposed method outperforms several of the previously proposed state-of-the-art approaches [41, 42, 51-55].
2. A method to autonomously delineate an in focus OOI from a larger in focus ROI using the dominant gradients and *k*-means segmentation.
3. An unsupervised tri-region classification method for LDOF images.
4. An automatic Gestalt-based technique for the categorisation and labelling of the objects and regions in a LDOF image.
5. A novel and improved unsupervised method for vanishing point (VP) detection. The straight edge detection and subsequent interrogation of the converging and intersecting dominant straight lines as well as the final extrapolation of the optimal VP is performed entirely in the image-plane coordinate space. The results show the proposed method either matches or outperforms several of the previously proposed approaches [23, 56-59].
6. An unsupervised method of sub-classifying a LDOF image based on the relative distance relationships between the OOI and non-OOI objects and regions.
7. An unsupervised method of extrapolating gradient-planes through the correlation of the extracted Gestalt-based ground region with VP estimation.
8. An autonomous depth map generation method for LDOF images. The results show the proposed method outperforms several of the previously proposed approaches [60-65].

9. An improvement to the DIBR autonomous hole-filling techniques used for the resolution of dis-occlusions and inconsistencies associated with 3D image warping. Results show the proposed approach outperforms other previously proposed methods [62, 65].

E. Structure of the Thesis

This research investigates three notable challenges associated with the overall automatic conversion of single 2D LDOF images into 3D.

The first research problem involves two sub-components. The first sub-component (stage 1 of the proposed approach), which deals with the extraction of the region-of-interest (ROI) from within a LDOF image, is discussed in **chapter 2**. The second sub-component (stages 2 and 3 of the proposed approach), which is concerned with the precise delineation and extraction of the object-of-interest (OOI) from within the ROI, is discussed in **chapter 3**. This is in support of the depth map generation process, whereby the accurate delineation of the OOI is necessary for improved accuracy. The title of **chapter 2** is "*Unsupervised Region-of-Interest Extraction based on the Correlation of Gradient and Higher-Order Statistics Saliencies*." The OOI within a LDOF image may either constitute the entire ROI or represent a subcomponent of a larger FG region. The problem with the latter is that without any a priori knowledge of the scene both the OOI and non-OOI region within the larger ROI become indistinguishable. The title of **chapter 3** is "*Unsupervised Matting of the Object-of-Interest in Low Depth-of-Field Images*."

The second research problem has three sub-components. The first sub-component involves the segmentation of the non-OOI region and the subsequent labelling of these objects and sub-regions. In addition to the OOI, a distinction and associated categorisation of the ground and sky regions as well as the individual objects on top of or within these regions needs to be made. The second sub-component is concerned with the allocation of depth to these specific regions. Although a LDOF may provide some distinction between the foreground and background, the relative depths of the objects within and across these regions are not so easily distinguishable. Furthermore, the allocation of depth to certain regions, such as the ground and the object on the periphery of the ground, are incremental while others may be static.

The first two sub-components are combined into a single research component (stages 4 and 5 of the proposed approach) and are discussed in **chapter 4**. This is entitled "*Autonomous Depth Map Generation of Single 2D Low Depth-of-Field Images using Defocus, Linear Perspective and Gestalt Principles*."

The third sub-component (stage 5 of the proposed approach), which involves the estimation of the vanishing point in the image, is discussed in **chapter 5**. This is necessary for determining the incremental depth of the associated gradient-plane needed for the assignment of relative depth

values to the objects and regions across the image. The title of **chapter 5** is "*Vanishing Point Detection of Man-Made Environments*."

The third research problem (stage 6 of the proposed approach), which involves the handling of the dis-occlusions associated with the 3D image warping component, is discussed in **chapter 6**. The resolving of these discontinuities has a direct bearing on the efficacy of the final 3D product. There is no one optimal resolution to this problem. Moreover, irrespective of how effective the hole-filling is for either one or several of the test images there is no actual way of quantifying or predicting future successes. This is owing to the fact that this data doesn't actually exist. The title of **chapter 6** is "*Stereoscopic Image Synthesis of Single 2D Low Depth-of-Field Images using Depth Image-Based Rendering*."

Stage 7 of the proposed approach, which involves the generation of an anaglyph image from the synthesised stereoscopic image pair, is discussed in **chapter 7**. This is entitled "*3D (Anaglyph) Image Generation*."

Finally, **chapter 8** concludes the thesis by summarising the key aspects of the research as well as discussing some future work.

II. UNSUPERVISED REGION-OF-INTEREST EXTRACTION BASED ON THE CORRELATION OF GRADIENT AND HIGHER-ORDER STATISTICS SALIENCIES

A. Introduction

The recognition and visual delineation of the region-of-interest (ROI) in an image plays a vital role in arenas such as multimedia entertainment, machine vision, depth estimation as well as 2D-to-3D conversion. There is no optimal solution for the unsupervised extraction of the ROI. However, by interrogating and correlating certain saliencies associated with low depth-of-field images the automatic delineation and subsequent extraction of the ROI may be possible.

Fig. 9 Depth-of-field [66]. (a) High DOF. (b) Low DOF.

In photography and cinematography, the term depth-of-field (DOF) refers to amount of focus (sharpness) or defocus (blurriness) present in an image. Image or video capture is broadly categorised into two DOF techniques. These include deep or high DOF and shallow or low DOF (LDOF). The LDOF technique tends to spotlight the subject by minimising or removing distractions through the defocussing or blurring of the background (BG). This provides hints as to where in the image or on the screen the viewers should be *focussing* their attention. Fig. 9 provides an illustration of both these DOF methods.

Moreover, an image captured by a camera using a standard convex lens with a LDOF setting will contain objects at a particular distance from the camera that are in focus, as well as objects that are out of focus, to varying degrees, depending on their distance in front of or behind the focal plane. These attributes may provide a means of both extrapolating the relative depths across an entire image as well as delineating the ROI. In LDOF images the sharp (focussed) regions are

usually associated with the foreground (FG) regions while the blurred (defocussed) regions are typically associated with the BG regions, as illustrated in Fig. 9(b).

Several methods have been proposed for FG or ROI extraction [41, 42, 51, 52, 55, 67, 68]. This research, which may be considered as an amalgamation and extension of the approaches proposed by Won et. al [68] and Kim [41], investigates and proposes a method for the detection and subsequent extraction of the ROI in LDOF images through the interrogation and correlation of high-frequency saliencies using gradient and higher-order statistics (HOS) analysis techniques together with cluster-based segmentation.

This paper is organised as follows: Section B provides a brief discussion of some of the related work. Section C provides a description of the proposed model as well as the necessary steps required to perform unsupervised ROI extraction. Section D reports the experimental results and provides are comparison of the proposed method against other state-of-the-art techniques. Section E provides a brief summary of the research and section F closes the paper.

B. Related Work

Two simple approaches were initially proposed for ROI extraction; these include direct edge analysis and region- or texture-based segmentation [69].

For the edge-based methods, their efficacy is directly related to the homogeneity of the grey-level intensities. This dependency is shown to be limiting [70], especially for natural objects, where boundary edges often remain disconnected and undefined. Even with the incorporation of edge-linking techniques these methods nevertheless provide insufficient information needed for effective ROI extrapolation.

The region-based techniques usually involve two processes. The first is detection, the purpose of which is to be able to locate the ROI. The second is extraction, the aim of which is to effectively delineate the closed boundaries of the ROI.

In LDOF images the subset of regions that constitute the complete ROI may be inhomogeneous when considering low-level features such as texture and colour. Another concern is that with unsupervised region segmentation no baseline criterion or reference exists, resulting in the problem becoming ill-posed [42].

As a consequence other more traditional segmentation methods such as *k*-means [71], mean shift [72], graph-cut [73-75], active contour [76, 77] and learning-based [78] segmentation may be unsuitable for the unsupervised delineation of the ROI in LDOF images. Fig. 10 provides an illustration of the segmentation results obtained through the application of some of these algorithms to several LDOF images. Some of the concerns associated with these aforementioned techniques include initialisation, inconsistent cluster sizing, iterations and long processing times.

Fig. 10 Common segmentation techniques applied to LDOF images. (a) Original image; (b) *k*-means segmentation [71] (cluster size = 4); (c) Mean shift segmentation [79] (supervised for cluster size = 4); (d) Graph-cut segmentation [75] (region/cluster size = 4).

Based on the experimental results illustrated in Fig. 10 as well as owing to the concerns and inconsistencies associated with these more traditional segmentation approaches other techniques have been proposed that deal more explicitly with the unsupervised segmentation of the ROI in LDOF images.

These methods are broadly classified into the three categories. The first is multi-resolution wavelet analysis [61, 62, 70, 80, 81], the second is Gaussian blur analysis [42, 51, 53, 60, 63, 69, 82-84], which includes depth from defocus (DfD), and the third is local variance and higher-order statistics (HOS) analysis [41, 65, 67, 68, 85]. The first and third methods are more closely associated with explicit ROI estimation while the second method deals more broadly with the estimation of overall relative depth.

When digitally transposed, the ROI in a LDOF image may be mathematically described as regions of high-frequency and the BG may be mathematically described as regions of low frequency.

Multi-resolution wavelet analysis techniques attempt to attenuate the low frequency components while exploiting the high-frequency components in an image as a result of energy compaction through subband decomposition. A focused region in an image will exhibit low attenuation of its high frequency components, while defocused regions will have a large attenuation of the high frequency components. This suggests that the measure of the image blur caused by the defocus is directly related to the amount of local spatial frequencies. Wavelet transforms [86] provide a means to analyse signal structures of different sizes and spatial frequencies. Each wavelet coefficient is spatially localised since only local segment features are included, and since sub-band decomposition allows for non-overlapping of frequency bands, each wavelet coefficient is also frequency localised [80].

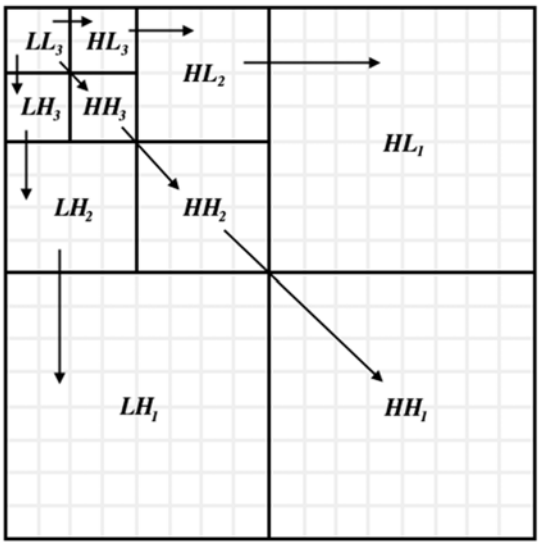


Fig. 11 Three-scale wavelet decomposition [80]

One method proposes the extraction of the ROI through analysis of the high frequency coefficients of the Haar, or short length filter Daubechies', wavelet transforms [70]. Another method proposes using a three scale 1D wavelet transform to generate the wavelet transform coefficients [62]; the sub-bands are illustrated in Fig. 11, where $LH_1 - LH_3$, $HL_1 - HL_3$ and $HH_1 - HH_3$ correspond to the high to low horizontal, vertical and diagonal frequency components respectively and LL_3 represents the DC component. D–D

The source image is initially divided into macro-blocks of 16×16 pixels. A three-scale wavelet transform is performed on each macro-block and each pixel in the macro-block is then assigned a value equal to the sum of the total non-zero wavelength coefficients (up to a maximum of $16 \cdot 16 = 256$). Finally a macro-block level depth map is directly extrapolated by interpreting each block as having a depth level of between 0 and 255 (either through normalisation or subtraction of 1). To refine the depth map the Lipschitz edges are considered.

The Lipschitz exponent, α, is described as a measure of the number of times a signal point is differentiable [81]: $\alpha = 1$ implies a the signal is differentiable once, $\alpha = 0$ implies the signal is a step edge, and $\alpha = -1$ implies the signal is a Dirac impulse. Using the modulus maxima of the wavelet coefficients across successive scales the Lipschitz exponent is calculated, whereby an edge is estimated to be in focus if $-1 < \alpha < 0$ and defocused in $0 < \alpha < 1$.

The final depth map is generated by integrating the Lipschitz exponents of the edges and the macro-block depth map into image rows and allocating each structure a distinct depth value according to each row subsection. The accuracy of the intial depth map is improved by performing local high frequency wavelet analysis of each pixel [61] as opposed to a block of pixels [62]; however, computation time is significantly increased. Futhermore, the initial edge map is improved by determining the Lipschitz exponents using 2-D wavelet analysis, as opposed to 1D wavelet analysis and enhanced by performing edge linking [62]. Finally, colour-based segementation is incorporated into the refinement of the final depth map.

In LDOF images some uniform FG regions may exhibit low frequency attributes and some busy textured BG regions may exhibit high-frequency attributes. Although it is reasonable to assume that a significant majority of these high-frequency regions represent the FG, these regions alone may not completely represent the ROI. Athough multi-resolution wavelet analysis techniques have their merits, a concern is that some of these inconsistencies are neglected. This is owing primarily to the energy compactation of the subband decomposition, which may result in the unavoidable discarding or attenuation of useful information pertaining to the ROI [42, 80].

Gaussian blur analysis techniques, such as DfD, deals with how objects are perceived depending on their distance from the camera's focal plane based on the relationship between the focused and defocused regions in an image [87]. In LDOF images objects will exhibit a certain amount of blur (out of focus) relative to their position very near to, or very far from, the object in

focus. For the purpose of ROI extraction it is not necessary to have a complete description of all the depths in an image. Nevertheless, ROI extraction may be considered as a subdivision of depth estimation, since in a typical LDOF captured image part, if not all, of the FG is considered to be nearest in distance from the camera. Direct Gaussian blur analysis is more adequately suited to deriving overall relative depth values. As a consequence, a large amount of unnecessary data may exist and needs to be accounted for and effectively discarded in order to provide adequate estimation of the ROI. Even with the incorporation of bilateral filtering some of the sporadic outlier regions may be misinterpreted as belonging to the FG and several of the intensity consistent and smooth FG regions may be misinterpreted as belonging to the BG. For a more expansive discussion on DfD and other Gaussian blur analysis techniques refer to *p*. 79.

The high-frequency components in a LDOF image may provide a foundation for the estimation of the ROI. However, exploiting only these attributes may result in inconsistencies occurring in both the defocused and focused elements of an image. This is owing to some sections of defocused regions having busy texture areas with high enough frequency components so as to be considered as in focus, and some sections of focused regions having constant texture areas with insufficiently high-frequency components so as not to be considered as in focus.

As a means of minimising the irregularities associated with multi-resolution wavelet and Gaussian blur analysis techniques, in terms of ROI extraction, a method is proposed whereby local variances (second-order moments) are considered as a means of extracting the energy of the high-frequency components in grayscale images through the application of block-wise maximum a posteriori (MAP) segmentation on every pixel [68]. This method is later expanded to include HOS, or more specifically fourth-order moments [41]. The interrogation of high-frequency components using HOS (third-order and above moments), when compared to local variances, is shown to provide an improvement to both Gaussian noise suppression as well as the preservation of some of the non-Gaussian information. The HOS method proposed by Kim [41] is subsequently modified to incorporate colour images [67, 85], as well as include additional refinements to increase performance of OOI segmentation using block-based methods [67]. The following discussion provides a mathematical description of the HOS technique proposed for ROI estimation.

Given an $M \times N$ input image I, the n^{th}-order moment is described as

$$\hat{m}^{(n)}(x,y) = \frac{1}{N_\eta} \sum_{(s,t)\epsilon\eta(x,y)} \left(I(s,t) - \hat{m}(x,y)\right)^n, \tag{1}$$

$$\hat{m}(x,y) = \frac{1}{N_\eta} \sum_{(s,t)\epsilon\eta(x,y)} I(s,t), \quad (0 \le x < M, 0 \le y < N),$$

where $\eta(x, y)$ is a set of neighbouring pixels centred at (x, y), $\hat{m}(x, y)$ is the sample mean of $I(x, y)$ and N_η is the size of η [67, 68]. For an RGB image the maximum n^{th}-order moment of each of the pixels across the three channels is considered, giving

$$HOS_{RGB}(x, y) = \max(\hat{m}_R^{(n)}(x, y), \hat{m}_G^{(n)}(x, y), \hat{m}_B^{(n)}(x, y)) \,. \tag{2}$$

In the case of the 4$^{\text{th}}$-order moment operation the resulting dynamic range of the *HOS* values may exceed 2×10^9. As a consequence a down-scaling factor (*DSF*) is introduced such that

$$DSF = \begin{cases} 300 & \text{for } \max\left(HOS_{RGB}(x, y)\right) < 10^7 \\ 700 & \text{for } \max\left(HOS_{RGB}(x, y)\right) < 10^8 \\ 1000 & \text{otherwise.} \end{cases} \tag{3}$$

In the case of the local variance (second-order moment) $DSF = 1$. Applying the *DSF*, the final saliency HOS map (Π_{HOS}) is then defined as

$$\Pi_{\text{HOS}}^{(n)}(x, y) = \min(255, \frac{HOS_{RGB}(x,y)}{DSF}) \,. \tag{4}$$

The HOS method is shown to be more robust than the local variance method, where the former is shown to yield denser values in the focused regions while suppressing noise in the defocused regions [41]. Moreover, the HOS technique is shown to be extremely effective in high-frequency component analysis and together with direct gradient analysis is applied to the problem of ROI detection and subsequent delineation.

C. Proposed Approach

LDOF images by definition contain regions that are in focus (focused) as well as regions that are out of focus (defocused). From observation, the focused regions have a higher probability of being associated with areas closer to the camera i.e. the foreground (FG) or the region-of-interest (ROI) in an image, while the defocused regions have a higher probability of being associated with areas further away from the camera i.e. the background (BG) in an image.

When digitally transposed it is understood that the FG in a LDOF image may be mathematically described as regions of high-frequency. This rudimentary understanding provides a foundation for the proposed model whereby the delineation of the ROI is achieved primarily through the interrogation of the high-frequency (HF) components. The HF attributes considered are edges, gradients and heuristics associated with higher-order statistics (HOS). In the proposed

model these three these attributes are correlated in order to form a more robust means of identifying and effectively isolating the closed boundaries of the ROI from the rest of the LDOF image. Moreover, as a means of refining the ROI, cluster-based segmentation and colour pixel analysis are also considered.

In this research the following process is proposed for the unsupervised extraction of the ROI from a single LDOF image:

The first step involves the extrapolation of three saliency maps. These include the edges in the image as well as the gradient and 4^{th}-order moment (HOS) of each pixel. Although an edge map is a binarised array it is nevertheless considered as another type of saliency map.

The second step involves the extrapolation of two separate approximations of the ROI by individually analysing the gradient and HOS saliency maps. It is assumed the ROI, or more specifically the boundaries of the FG regions, will be associated with higher absolute values in the gradient and HOS saliency maps and vice versa for BG regions. Although some BG regions may contain higher values than some FG regions, it is reasonable to assume that the majority of the higher values will be associated with the FG regions. Since the gradient and HOS saliency maps contain all the derived attribute values for all pixel locations in the image, two separate approximations of the ROI are produced by statistically thresholding these maps. These ROI estimate maps are expected to contain pixel points belonging to both the FG and BG regions. However, a significant majority of the pixels are expected to belong to the FG.

The third step involves the extrapolation of a baseline reference for the ROI. Owing to the ill-posed nature of the problem there is no a priori reference for the ROI. To resolve this quandary a baseline reference map or mask is proposed. This ROI baseline reference mask is expected to contain the locations of pixels having the highest probability of representing the edges and/or clusters of FG regions. The ROI estimate maps may contain a modicum of BG outliers. However, the ROI baseline reference mask is expected, in principle, to contain relatively few to no edges or regions belonging to the BG. Moreover, if the image contains an OOI, then the majority of the prominent connected components in this baseline mask is expected to represent some part (most likely the boundary regions) of the OOI. This ROI baseline reference mask is extrapolated through a more adaptive or refined statistical thresholding of the gradient saliency map.

The fourth step involves the extraction of the ROI through the correlation of the baseline ROI reference mask, gradient ROI estimate map, HOS ROI estimate map and edge saliency map. The fifth and final step involves the refinement of the ROI using *k*-means segmentation as well as CIEDE2000 colour-difference analysis.

1) Saliency Maps

i. Edge Saliencies

The edge saliency map (Π_{EDGE}) is produced through the merging of six extrapolated edge maps. The first four edge maps are derived through the application of the Sobel edge detection method to the R, G and B layers, as well as the grayscale version, of the image. The fifth and sixth edge map are based on the application of both the Sobel and Canny edge detection methods to the lightness (L) layer of the transformed CIEL*a*b* version of the original colour image. Application of edge detection to the a* and b* layers of the CIEL*a*b* colour space is not considered owing to the increase in the number of inconsistencies as well as spurious and branchpoint edges. The proposed method for Π_{EDGE} is shown, from experiment, to be the most robust when dealing with LDOF images.

Edge maps are binarised arrays that provide an estimate of the location of edges, around and within objects and regions, across an entire image and while this is useful in some regards it is usually insufficient in providing information that may be used to adequately discern between the FG and BG regions of an image. In the proposed algorithm Π_{EDGE} is not considered as a direct means of ROI estimation. However, it is employed in a supplementary capacity to correlate the gradient- and HOS-based approximations of the ROI.

ii. Gradient Saliencies

The direct analysis of gradients in an image is usually associated with edge detection rather than focus/defocus analysis. However, both these arenas are considered to be somewhat related. It is understood that for LDOF images there is a higher probability of the focused FG regions (especially edges) having local neighbourhood gradients noticeably greater than the defocused BG regions. This is owing, in the main, to the point spread nature of defocused light, which will typically exhibit gradients with lower values. As such, in this research, the extrapolation of the gradients across the image is considered to be a rudimentary form of blur analysis.

There currently exist several techniques for estimating gradients in images. From experiment, the Sobel gradient method is shown to be the most robust for LDOF images. A typical gradient map will consist of gradient magnitude as well as associated direction values. However, for the generation of the gradient saliency map, denoted by Π_{GRAD}, the direction values are not considered. Results of Π_{GRAD} are illustrated in Fig. 12(b). The more pronounced components are indicative of the high gradient FG regions. For display purposes the pixel intensities are scaled across 7 grey levels as opposed to 256.

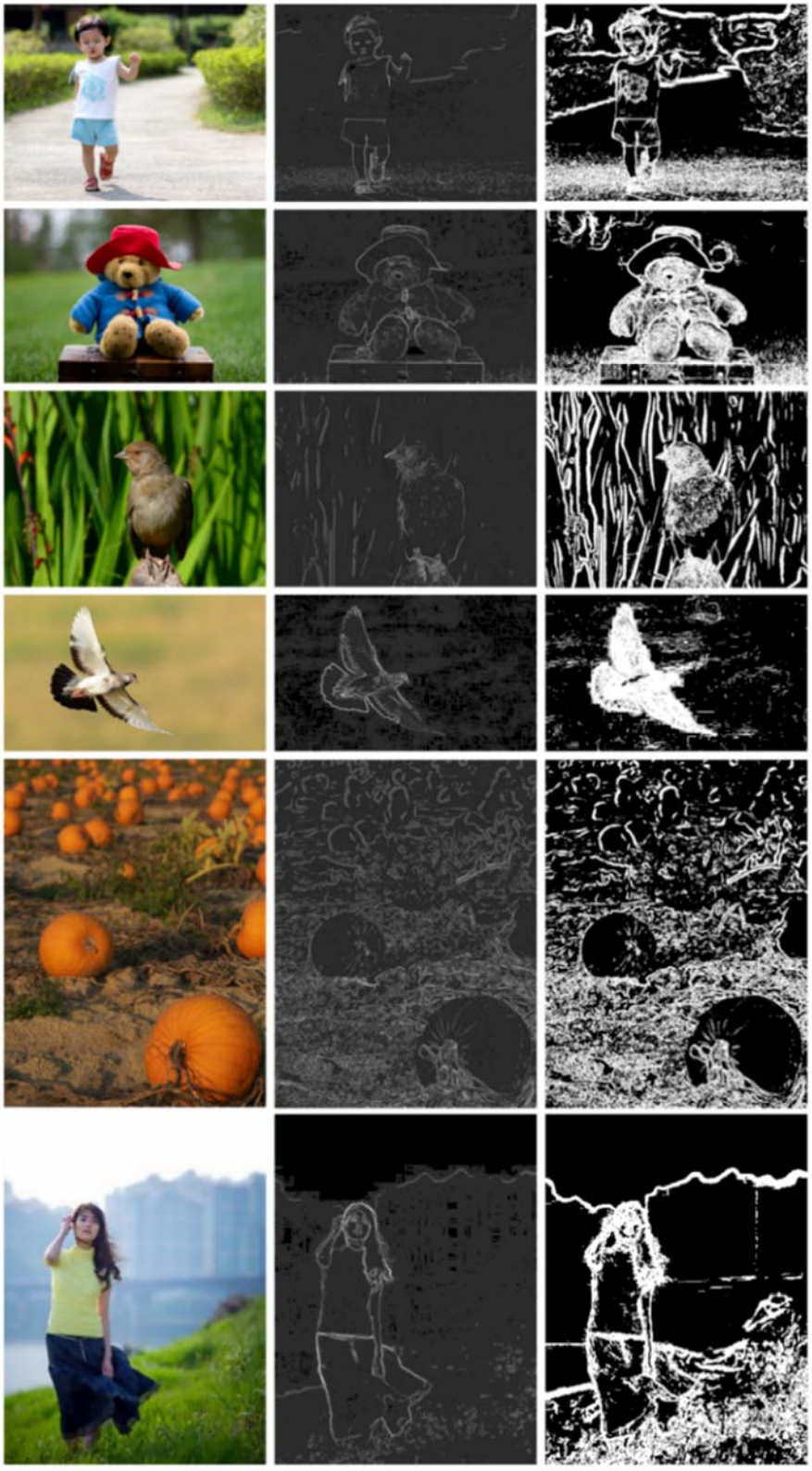


Fig. 12 Gradient saliencies. (a) Original images; (b) Full gradient maps; (c) Gradient-based ROI estimate maps.

From experiment, an approximation of the ROI is produced by applying a threshold to Π_{GRAD}. This is referred to as the gradient-based ROI estimate and is denoted by Π_{GRAD}^{ROI}. The threshold is determined by extracting the top 30% of the nonzero gradient values in Π_{GRAD} and using the lowest gradient value from this selection. The reason for adopting this approach is that the gradient values are relative and may often be clustered above or below a threshold. As a consequence, a percentage-based threshold may therefore be considered as being more effective than selecting all pixels whose value is above a certain threshold, like in the case of edge detection. Results of Π_{GRAD}^{ROI} are illustrated in Fig. 12(c).

iii. HOS Saliencies

The interrogation of only the high-frequency components, as in the case of afore-discussed gradient saliencies, may result in inconsistencies occurring in both the focused and defocused elements of an image. As a means of minimising these irregularities as well as to provide an additional correlator for these elements, the proposed model also considers the heuristics associated with the higher-order statistics (HOS) analysis of the image. In the proposed model the fourth-order moment HOS algorithm proposed by Kim [41] and Kim et. al [67] is used to derive the HOS saliency map (Π_{HOS}). Results of Π_{HOS} are illustrated in Fig. 13(b). The more pronounced components are indicative of the FG regions. For display purposes the pixel intensities are scaled across 5 grey levels as opposed to 256.

In the proposed model an approximation of the ROI is produced by applying a threshold to Π_{HOS}. This is referred to as the HOS-based ROI estimate and is denoted by Π_{HOS}^{ROI}. From experiment the threshold is chosen as the lowest value between 0.5 and the mean of all the values in Π_{HOS} between 0.3 and 1. Results of Π_{HOS}^{ROI} are illustrated in Fig. 13(c).

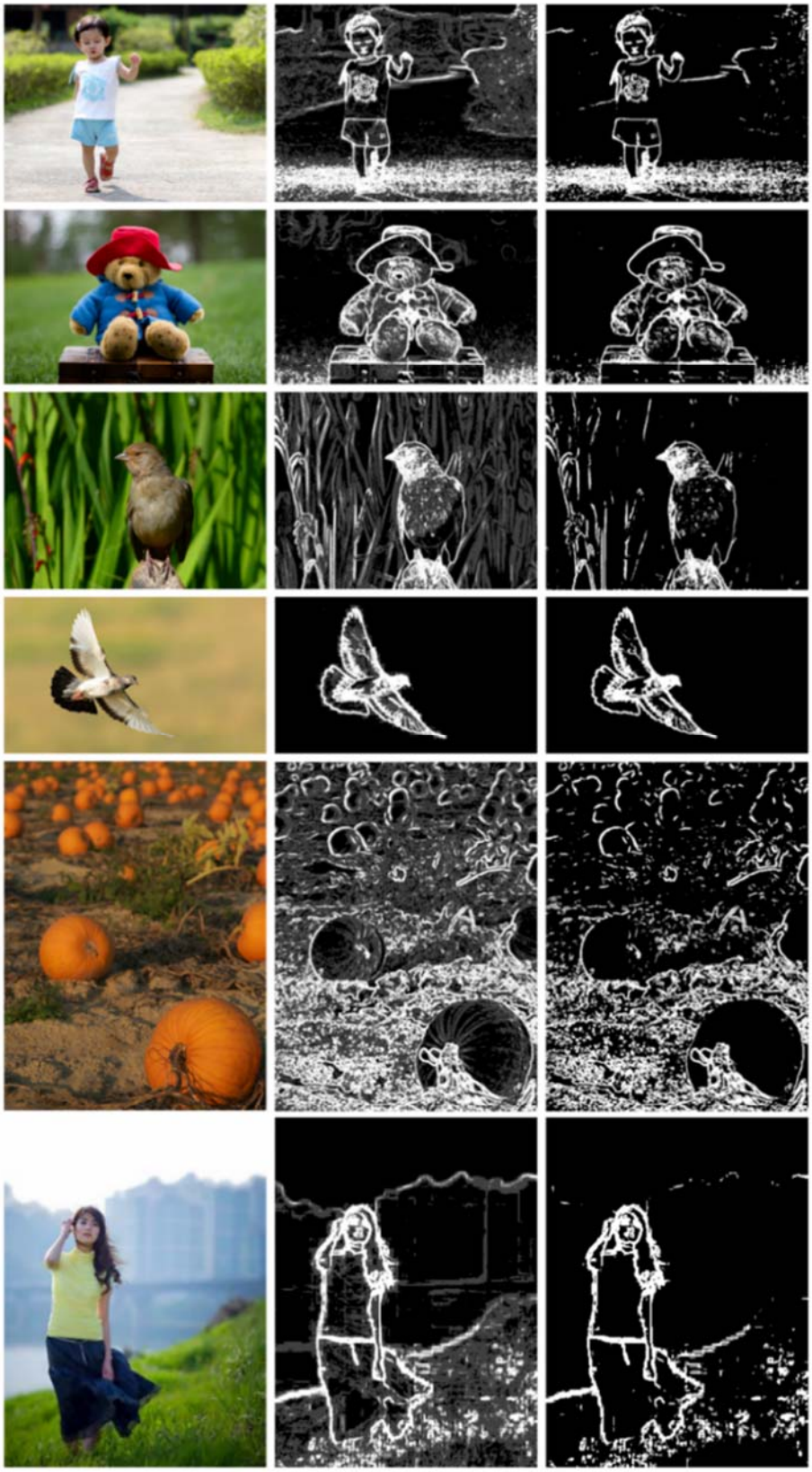


Fig. 13 HOS saliencies. (a) Original images; (b) Full HOS maps; (c) HOS-based ROI estimate maps.

2) ROI Detection and Extraction

i. ROI Baseline Reference Mask

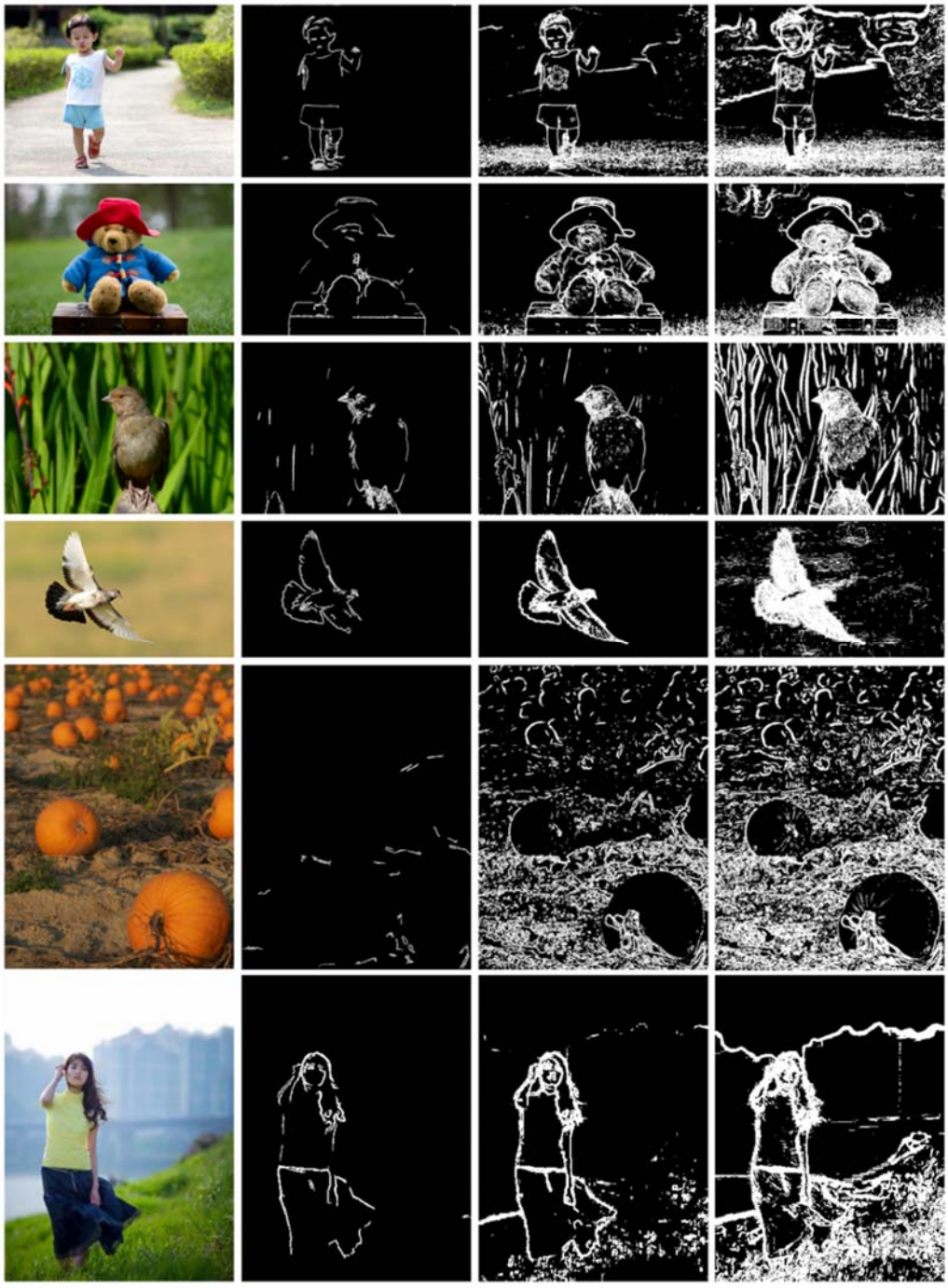


Fig. 14 ROI reference masks. (a) Original images; (b) ROI baseline reference; (c) Gradient-HOS intersection; (d) Gradient-HOS union.

It is necessary to have some sort of rudimentary ground-truth reference for the ROI in order to correlate the data from the saliency maps. However, the problem is ill-posed. To resolve this conundrum, it is shown from experiment that by applying a more restrictive threshold to the gradient saliency map Π_{GRAD} a robust baseline reference for the ROI may be extrapolated. The threshold is determined by considering the top 3% of pixels from the nonzero gradient values in Π_{GRAD} and using the lowest gradient value from this selection. Subsequently any connected component having a pixel count less than 0.015% of the area of the image is removed. Results of the ROI baseline reference mask, denoted by Π_{ROI}^{REF}, are illustrated in Fig. 14(b).

From Fig. 12 and Fig. 13 it may be seen that the estimate maps Π_{GRAD}^{ROI} and Π_{HOS}^{ROI} contain both FG as well as BG regions while Π_{ROI}^{REF} provides a more refined description of the boundaries of the ROI and therefore may be considered as a form of ROI detection.

ii. ROI Estimation

From experiment the optimal estimation of the ROI is achieved through the correlation of the intersection as well as union of the gradient and HOS ROI estimate maps (Π_{GRAD}^{ROI} and Π_{HOS}^{ROI}) together with the edge saliency map Π_{EDGE} and the ROI baseline reference mask Π_{ROI}^{REF}.

A first approximation of the ROI is based on the intersection between Π_{GRAD}^{ROI} and Π_{HOS}^{ROI}. This is illustrated in Fig. 14(c) and denoted by Γ_1. Expansion and refinement of this region is performed using morphology. Regions and edges that are within a close pixel proximity of each other are assumed to be a subset of a larger connected region. In the proposed model, for the intersection approximation, this proximity is chosen to be 2 pixel spaces. Initially the region is expanded by performing dilation, filling and erosion. Subsequently, the region is refined by firstly eroding the connected components and secondly, removing any cluster not associated with Π_{ROI}^{REF} and thirdly, dilating the resulting connected components. Finally, the ROI is expanded by incorporating edges together with morphology. Initially the region is expanded by including Π_{EDGE}. This is followed by dilation, filling and erosion. Subsequently small connected components and lines are removed and any of the newly expanded regions not associated with Π_{ROI}^{REF} are excluded. This expansion of Γ_1 is denoted by Γ_1^{ROI}. From experiment, in order to account for the proposed 2 pixel proximity association, a 3×3 square structuring element is shown to be adequate for the dilation and erosion. This is illustrated in Fig. 15(b).

Fig. 15 Initial ROI Estimation. (a) Original images; (b) ROI approximation 1; (c) ROI approximation 2; (d) ROI Correlated ROI estimation.

A second approximation of the FG is based in the union between Π_{GRAD}^{ROI} and Π_{HOS}^{ROI}. This is illustrated in Fig. 14(d) and denoted by Γ_2. Initially this region is refined by removing any cluster not associated with Π_{ROI}^{REF}. Subsequently, the ROI is expanded by incorporating edges together with morphology. As with the previous scenario, regions and edges that are within a close pixel proximity of each other are assumed to be a subset of a larger connected region. In the proposed model, for the union approximation, this proximity is chosen to be 1 pixel space.

Initially the region is expanded by including Π_{EDGE}. This is followed by dilation, filling and erosion. Small connected components and lines are removed. Any of the newly expanded regions not intersecting or bordering with Γ_1 are excluded. This expansion of Γ_2 is denoted by Γ_2^{ROI}. From experiment, in order to account for the proposed 1 pixel proximity association, a 2×2 square structuring element is used for the dilation and erosion. This is illustrated in Fig. 15(c).

It is assumed that Γ_2^{ROI} contains both FG and BG elements. However, from experiment it is observed that a significant portion of the BG is actually excluded while essentially retaining the entire ROI. This is a crucial outcome of the proposed model. Based on this understanding the ROI is estimated by expanding and refining Γ_1^{ROI} using Γ_2^{ROI}. This is illustrated in Fig. 15(d).

Firstly, Γ_1^{ROI} is refined by performing an intersection with Γ_2^{ROI}. Secondly, all regions in Γ_2^{ROI} intersecting with Γ_1^{ROI} are set to 0. Thirdly, the remaining regions in Γ_2^{ROI} are separated using the edge saliency map Π_{EDGE}. Fourthly, from experiment, if more than 50% of the perimeter of the applicable connected component is adjacent to Γ_1^{ROI} then it is considered to be a valid extension of the ROI. Further refinements of the ROI are performed by evaluating the outer regions of the ROI through edge delineation.

The key benefit of the proposed method is that an extremely robust and accurate means of detecting and delineating the closed boundaries of the ROI in a LDOF image is achievable by only considering the correlation of Π_{GRAD}^{ROI}, Π_{HOS}^{ROI}, Π_{EDGE} and Π_{ROI}^{REF}.

iii. Final ROI Extraction

The method presented in the previous section provides an extremely adequate estimation of the ROI. For improved precision of the ROI, specifically around the border regions, further refinements are considered using segmentation together with colour difference analysis.

Over the past half century there have been a significant number of proposals developed for image segmentation [43, 88-91]. However, there is no one particular algorithm that satisfies all possible types of scenarios. Segmentation techniques are broadly divided into three categories [92]. The first is edged based, the second is region-based and the third is special theory based. The first category involves methods based on identifying abrupt changes in the intensity of image pixels. These include histogram [93, 94] and gradient based [95, 96] techniques. The latter

two categories are based on identifying similarities in the intensity of image pixels. These include thresholding [90, 97], region-based [98-101] and clustering [102-104] techniques. In this research the *k*-means cluster-based segmentation technique is considered.

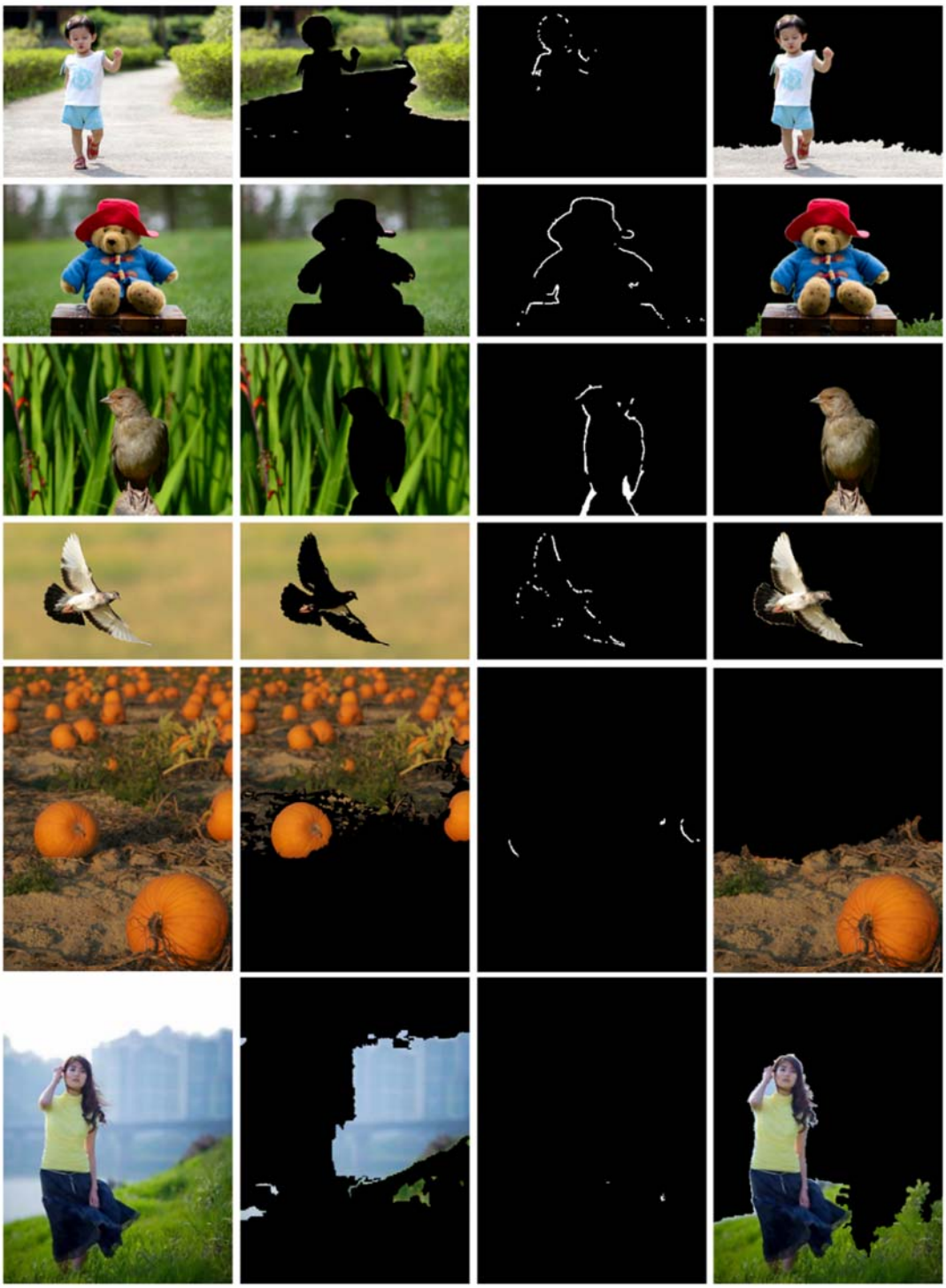


Fig. 16 Final ROI Extraction. (a) Original images; (b) Ground mask; (c) Invalid border regions; (d) Final delineated ROI.

The *k*-means clustering method, as described by Macqueen (1967), is among the most popular of the approaches employed in image segmentation [102]. An algorithm was initially proposed by Hartigan (1975) and subsequently modified for improved efficiency by Hartigan and Wong (1979) [71]. Gan et. al [105] provide the pseudo-code of the *k*-means algorithm employed in the proposed model. The distance measure is based on the Euclidean distance. The *k*-means algorithm is simple and fast and efficiently provides for rapid convergence to local minima of the error function.

The disadvantages of the *k*-means algorithm include increased sensitivity to initialisation as well as the need for pre-allocation of the number of clusters, where a wrong value may result in suboptimal results. Moreover, the algorithm is limited to spherical (or elliptical) shaping and is sensitive to outliers. Since no a priori data analysis is available the cluster loci are randomly initialised. A primary concern with this is that inconsistencies may occur when attempting to reproduce the segmentation results. A secondary concern is the selecting of the number of clusters and iterations, since unsupervised clustering may result in suboptimal over- or under-segmentation

Even though shortfalls may exist with *k*-means segmentation, it is nevertheless computationally advantageous and, from experiment, more suitable to LDOF images.

An initial unsupervised segmentation of the LDOF source image is performed using *k*-means clustering. To improve the consistency of the algorithm the initial cluster intensity locus value is pre-assigned to 128 instead of being randomly chosen. Moreover, to avoid or minimise the probability of local minima the algorithm is iterated through 3 cycles. The optimal outcome is subsequently chosen as the iteration with the largest distance between clusters in combination with the smallest distance within the clusters.

From experiment the number of clusters, n, is determined by comparing cluster collections for $n = m$ and $n = m - 1$ starting with $m = 7$ and decrementing until $m = 3$. Firstly, the percentage overlap of each of the arrays in cluster collection $n = m$ with respect to the ROI (determined in the previous section) is calculated and the maximum percentage overlap, denoted by $P_{\max}^{m}$, is extracted. Secondly, the percentage overlap of each of the arrays in cluster collection $n = m - 1$ with respect to the ROI is calculated and the maximum percentage overlap, denoted by $P_{\max}^{m-1}$, is extracted. Thirdly, if $P_{\max}^{m} > 0.85 P_{\max}^{m-1}$, then $n = m$. However, if $P_{\max}^{m} < 0.85 P_{\max}^{m-1}$, then $m = m - 1$ and the process is repeated. From experiment, the ceiling value of 7 and floor value of 3 is shown to lower the probability of over- and under-segmentation, respectively, and the limiting factor of 85% of the previous $P_{\max}^{n}$ is shown to adequately offset the loss of integrity of the ROI against the overall improved delineation.

An estimation of the BG of the image is initially derived using the segmentation clusters. This is illustrated in Fig. 16(b). From experiment the following process is employed for this purpose. Firstly, an entire cluster array is considered to belong to the BG if it has less than a 5% overlap

with the ROI. Secondly, the remaining FG cluster arrays are analysed by interrogating the individual connected components within the cluster array. In this case any connected component having less than a 10% overlap with the ROI is also considered as belonging to the background. Thirdly, any cluster array or connected component excluded from the BG are subsequently collated and reassessed. This is achieved by performing segmentation on the collated BG region using *k*-means clustering. From experiment a cluster value of 3 is chosen. The first two stages above are subsequently applied to these sub-segmented cluster arrays. To account for the reproducible inconsistencies associated with *k*-means segmentation the entire process above is iterated 3 times with the BG being constructed in an accumulative manner. Finally, any segment in the BG not located on the border region of the ROI or completely enveloped by the ROI is excluded.

An initial estimate of the invalid border regions is extrapolated by considering the intersection between the above derived segmentation-based BG region and the ROI. This is illustrated in Fig. 16(c). Subsequently, the applicable connected components are interrogated and validated using colour difference analysis.

There exist several methods for estimating the differences in the colour intensity of pixels [106]. An early attempt, referred to as the CIE76 colour-difference formula, was proposed in 1976. In this formula the difference in the colour of two pixels is estimated by analysing the pixels in the CIEL*a*b* colour space. To resolve some non-uniformities associated with the CIE76 colour-difference formula, an updated algorithm was proposed in 2000 by the members of the CIE Technical Committee 1–47, referred to as the CIEDE2000 colour-difference formula [107]. Although the CIEDE2000 colour-difference formula also analyses the pixels in the CIEL*a*b* colour space, it more specifically considers the CIEL*C*h* colour space. The latter refers to the lightness, chroma and hue, respectively, which may be calculated directly from the L*a*b* coordinates. Even though the CIEDE2000 colour-difference method is substantially more complicated and computationally involved than its early predecessor, the CIE76 colour-difference formula, it nevertheless provides a significantly more accurate means of comparing minor differences in colour intensity between two pixels or clusters of pixels [108].

In the proposed model the segmented-based invalid regions are assessed by interrogating the perimeter pixels adjacent to the ROI using the CIEDE2000 colour-difference algorithm proposed by Sharma et al. [109]. If more than 60% of the adjacent pixels contain a colour-difference greater than 3 then the connected component is assumed to be an invalid ROI segment and subsequently excluded from the ROI. Owing to border transitions the pixels are compared up to an adjacent distance of 3 pixels. The perimeter pixels adjacent to the BG pixels are not considered. Results of the segment-based refinement of the ROI are illustrated in Fig. 16(d). This concludes the method proposed for the unsupervised extraction of the ROI from a single 2D LDOF image.

D. Results and Discussion

The proposed method is evaluated using a common database of a 121 LDOF images together with their ground-truth masks obtained from Li and Ngan (2007) [42] and the COREL dataset [110, 111]. A training set of a 160 LDOF images together with their ground-truth masks (excluding the 121 test images) were obtained from Li and Ngan (2007) [42], the COREL dataset [110, 111] as well as the World Wide Web. Subjective as well as objective comparative analyses are performed against seven other proposed methods [41, 42, 51-55]. Moreover, the precision and F-measure is also evaluated against the other proposed approaches.

In this section the ROI is denoted by Ω_{ROI} and the nonzero pixels associated with the ground-truth mask of the image are denoted by Ω_{GT}. In addition, the terms true positive, false positive and false negative are used as comparative descriptors. True positive denotes the number of pixels common to both Ω_{ROI} and Ω_{GT}. False positive indicates the number of nonzero pixels belonging to Ω_{ROI} but not to Ω_{GT}. False negative indicates the number of nonzero pixels belonging to Ω_{GT} but not to Ω_{ROI}.

Some of the representative results obtained by different methods are illustrated in Fig. 17. The original images in Fig. 17(a) are titled (from top to bottom) as *Dog*, *Bee-Flower-1*, *Owl*, *Strawberry*, *Beckham* and *Flower-Yellow*, respectively. Fig. 17(g) provides the corresponding ground-truths where the ROI is indicated by the white region and the BG is indicated by the black regions. Fig. 17(b)-(d) show the results obtained from the methods proposed by Kim (2005) [41], Li and Ngan (2007) [42] and Chen and Li (2012) [51], respectively. Fig. 17(e)-(f) present the results of the proposed method as well as the associated binarised formats needed for ground-truth analysis.

When considering Kim's method, although the six images exhibit minimal false positive regions, all of them exhibit some degree of BG misclassification, i.e. false negative regions, such as the holes in the *Owl* and *Dog* images as well as the missing stem in the *Flower-Yellow* image. In the case of Li and Ngan's proposal, several of these associated false negative regions are resolved However, some false positive regions are introduced, such as the green regions between the flower petals in the *Bee-Flower-1* image and the green region next to the owl's left leg in the *Owl* image. In comparison, while Chen and Li's method eliminates false positives such as the green regions between the flower petals in the *Bee-Flower-1* image it also misclassifies part of the bee's wing and the stem of the flower in the *Bee-Flower-1* image as belonging to the background as well as the right part of the flower bud in the *Flower-Yellow* image. Although there are some minor false positives in the proposed method, it is nevertheless comparable to Chen and Li's method and shows a noticeable improvement to Kim's and Li and Ngan's methods. Furthermore, in comparison, the proposed method also produces more complete ROIs as well as clearer boundaries for a significant number of the 117 test images (Refer to Fig. 20).

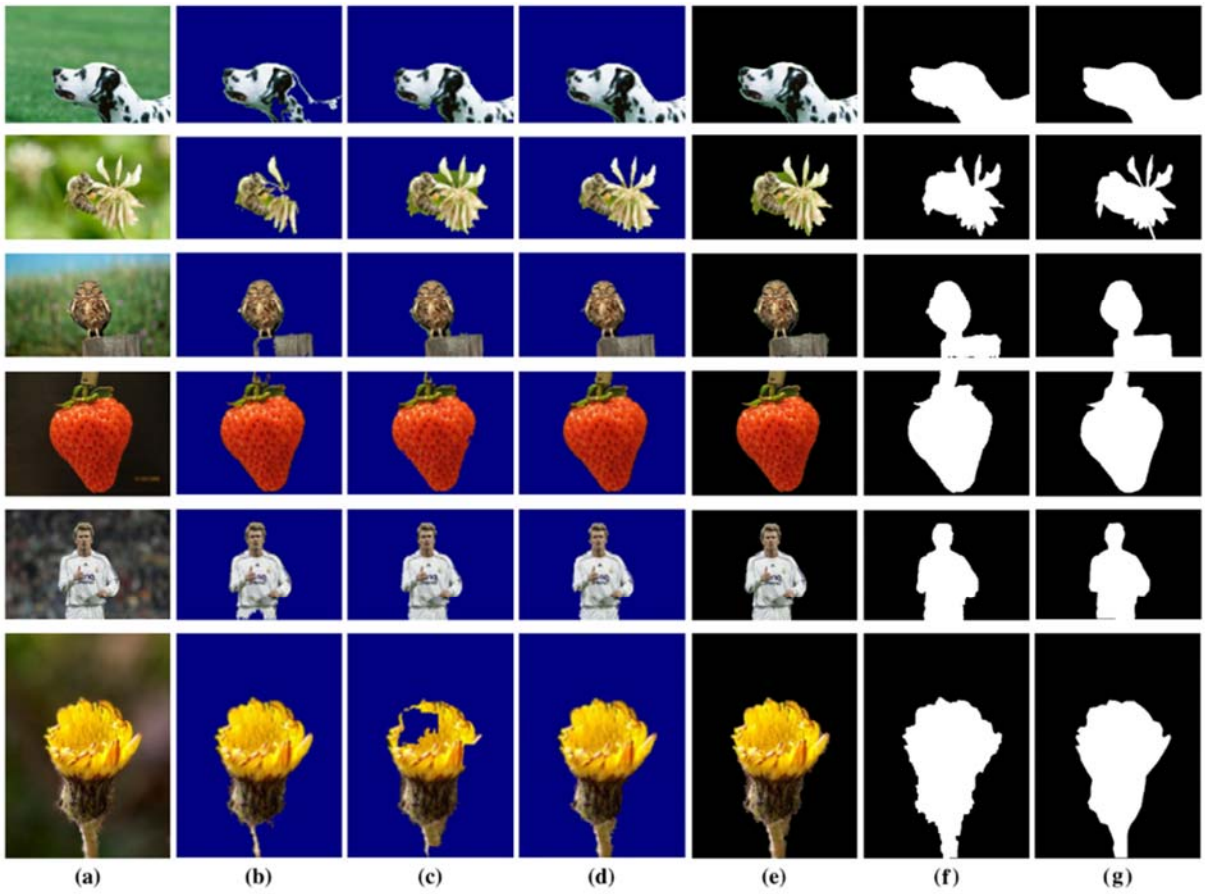


Fig. 17 Experimental results for test images: *Dog*, *Bee-Flower-1*, *Owl*, *Strawberry*, *Beckham* and *Flower-Yellow* (from top to bottom). (a) Original images; (b) Results from Kim (2005) [41]; (c) Results from Li and Ngan (2007) [42]; (d) Results from Chen and Li (2012) [51]; (e) Results of proposed method; (f) Binarised results of proposed method; (g) Binarised ROI ground-truths for original source image.

A second set of comparative results is illustrated in Fig. 18. The original images in Fig. 18(a) are titled (from top to bottom) as *Bee-Flower-2*, *Meerkat*, *Dragonfly*, *Flower-Red*, *Frog* and *Referee*, respectively. Fig. 18(g) provides the corresponding ground-truths where the ROI is indicated by the white region and the BG is indicated by the black regions. Fig. 18(b)-(d) show the results obtained from the methods proposed by Kim (2005) [41], Li and Ngan (2007) [42] and Li and Ngan (2011) [52], respectively. Fig. 18(e)-(f) presents the results of the proposed method as well as the associated binarised formats needed for ground-truth analysis.

When considering Kim's method, although the six images exhibit minimal false positive regions, all of them exhibit some degree of false negative regions, such as the stem in the *Bee-Flower-2* image, the missing heads in the *Meerkat* image and the holes in the *Dragonfly* and *Referee* images as well as part of the flower in the *Frog* image. In the case of Li and Ngan's (2005) proposal, several of these associated false negative regions are resolved with exception of the hovering bird in the *Meerkat* image, the missing tailpiece in the *Dragonfly* image

and the missing arm in the *Referee* image. However, some false positives are also introduced, such as the green regions around the outside of the flower in *Flower-Red* image.

In comparison, while Li and Ngan's (2011) supervised method eliminates most of these shortfalls, some false negative regions are introduced, such as the spine region of the adult meerkat in the *Meerkat* image. Although there exist some minor false positives in the proposed method, it is nevertheless comparable to Li and Ngan's (2007) unsupervised method as well as Li and Ngan's (2011) supervised method and shows a noticeable improvement to Kim's (2005) method. Furthermore, in comparison, the proposed method also produces more complete ROIs as well as clearer boundaries for a significant number of the 117 test images (Refer to Fig. 20).

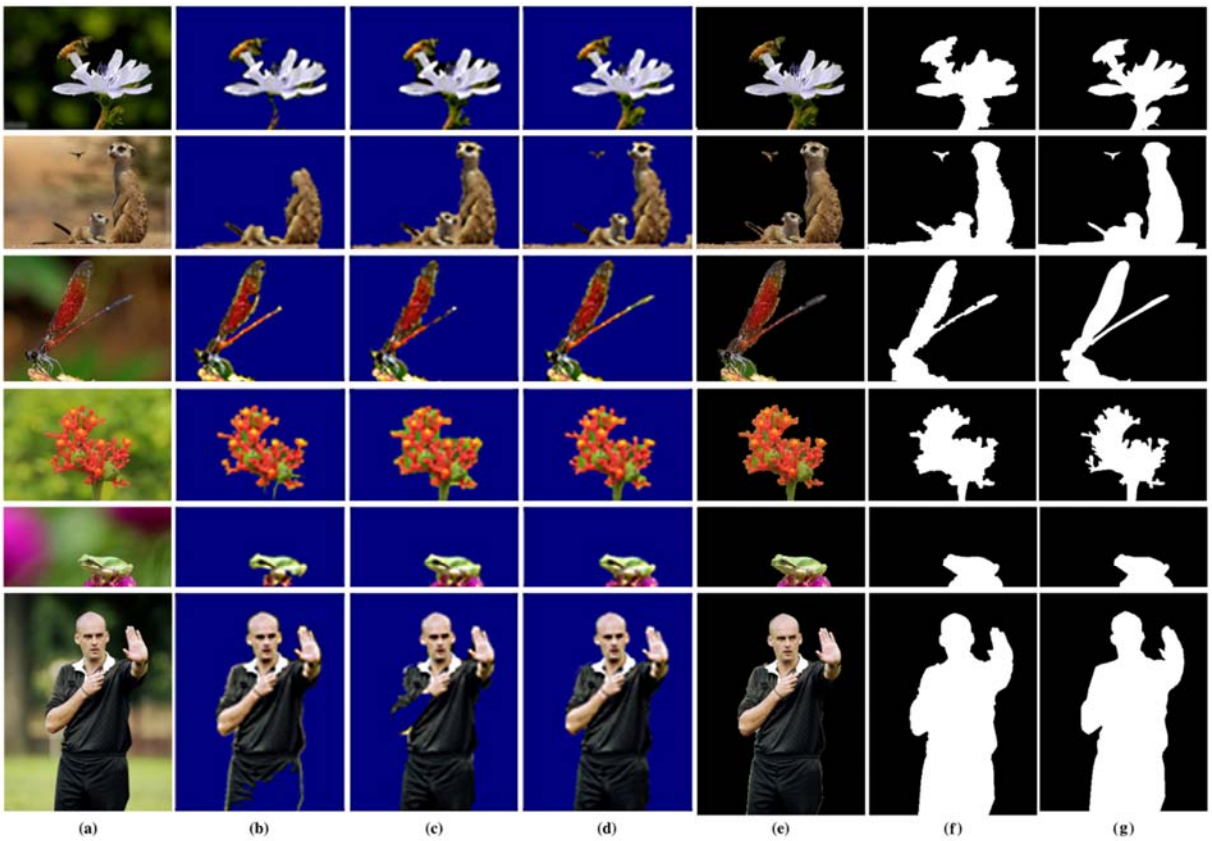


Fig. 18 Experimental results for test images: *Bee-Flower-2*, *Meerkat*, *Dragonfly*, *Flower-Red*, *Frog* and *Referee* (from top to bottom). (a) Original images; (b) Results from Kim (2005) [41]; (c) Results from Li and Ngan (2007) [42]; (d) Results from Li and Ngan (2011) – supervised approach [52]; (e) Results of proposed method; (f) Binarised results of proposed method; (g) Binarised ROI ground-truths.

A third set of comparative results taken from the COREL dataset is illustrated in Fig. 19. The original images in Fig. 19(a) are titled (from top to bottom) as *Moth*, *Cheetah*, *Hornbill* and *Soccer*, respectively. Fig. 19(b) provides the corresponding ground-truths. Fig. 19(c)-(g) show the results obtained from the methods proposed by Ye and Lu (2002) [54], Kim (2005) [41], Li and

Ngan (2007) [42], Liu et. al (2010) [53] and Rafiee (2013) [55], respectively. Fig. 19(h) presents the results of the proposed method.

When considering Ye and Lu's method, some distinct false positive regions are present, such as the background stem in *Moth* image and the ground region to the right of the cheetah in the *Cheetah* image. There also exist some false negative regions, such as the under-tail coverts and vent part of the hornbill in the *Hornbill* image as well as the leg of the soccer player to the left in the *Soccer* image. In the case of Kim's proposal, several of these associated false positive regions are resolved. However, some false negatives are introduced, such as the ground region to the left of the cheetah in the *Cheetah* image. The false negative in the *Soccer* image is resolved. However, the *Hornbill* image remains problematic. Li and Ngan's method exhibits similar discrepancies to those associated with Ye and Lu's method. However, some of the false positive regions are attenuated or eliminated altogether. In comparison, the method proposed by Liu et. al provides a consistent correlation with the provided ground-truth images. However, some false negatives are introduced, such as the moth's antennae in the *Moth* image and the top of the cheetah's head in the *Cheetah* image. Moreover, the under-tail coverts and vent part of the hornbill in the *Hornbill* image is still unresolved. Also, in comparison, the method proposed by Rafiee provides a consistent correlation with the provided ground-truth images. However, some false positives are introduced, such as the region between the ball and the head of the soccer player. Moreover, the under-tail coverts and vent part of the hornbill in the *Hornbill* image is still unresolved.

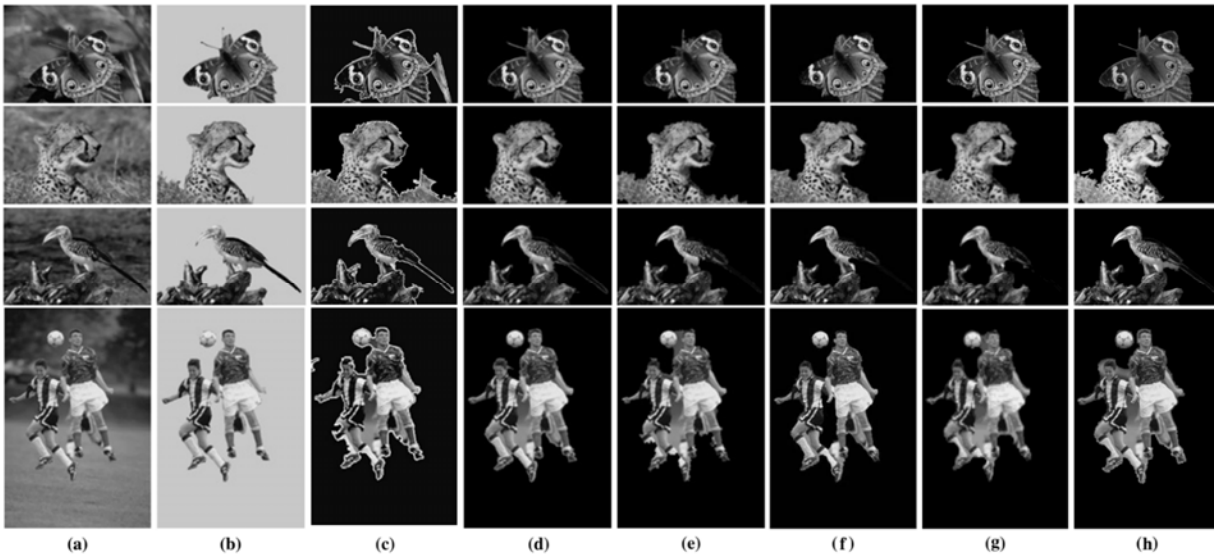


Fig. 19 Experimental results for test images from the COREL dataset: *Moth*, *Cheetah*, *Hornbill* and *Soccer* (from top to bottom). (a) Original images; (b) ROI ground-truths for original source images; (c) Results from Ye and Lu (2002) [54]; (d) Results from Kim (2005) [41]; (e) Results from Li and Ngan (2007) [42]; (f) Results from Liu et. al (2010) [53]; Results from Rafiee (2013) [55]; (h) Results of proposed method.

Although some minor false positives and false negatives exist in the proposed method, it is nevertheless comparable to the methods proposed by Kim and Liu et. al and Rafiee. Furthermore, a noticeable improvement is shown in comparison to Ye and Lu's as well as Li and Ngan's methods.

To provide an objective comparison of the proposed method against the aforementioned methods, the segmentation errors are computed across all of the 121 test images (117 from Li and Ngan [42] plus 4 from the COREL dataset [110, 111]) using the measurement criterion adopted by these investigations.

The segmentation error, which may be defined as the proportion of the misclassified pixels to the nonzero ground-truth pixels, is mathematically expressed by the following operation:

$$error = \frac{|\Omega_{ROI} \bigcup \Omega_{GT}| - |\Omega_{ROI} \bigcap \Omega_{GT}|}{|\Omega_{GT}|} \tag{5}$$

where Ω_{ROI} denotes the ROI, Ω_{GT} denotes the nonzero pixels associated with the ground-truth mask of the image and $|\cdot|$ indicates set cardinality.

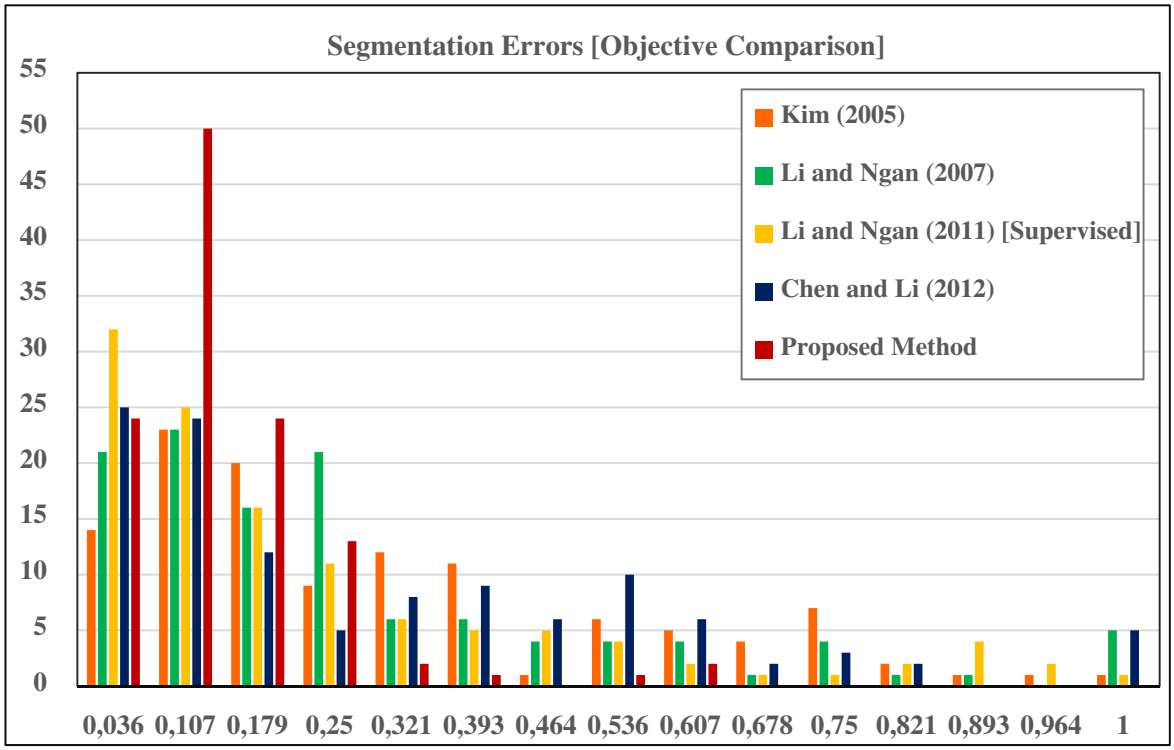


Fig. 20 Objective comparison. Segmentation errors of the 117 test images for Kim (2005) [41], Li and Ngan (2007) [42], Li and Ngan (2011) – supervised approach [52], Chen and Li (2012) [51] and the proposed method.

The distribution of the segmentation errors across the 117 test images [42] for Kim (2005) [41], Li and Ngan (2007) [42], Li and Ngan (2011) [52], Chen and Li (2012) [51] and the proposed method are illustrated in Fig. 20. Li and Ngan's (2011) approach is supervised compared to the rest of the approaches, including the proposed method, which are unsupervised. The results of the proposed method are obtained by taking the average error from 50 iterations of the proposed algorithm across the 117 test images.

Although the proposed unsupervised method shows less accuracy in the 0.036 range compared to Li and Ngan's (2011) supervised method and Chen and Li's (2012) unsupervised method, it nevertheless exhibits twice the accuracy of the next closest method (Li and Ngan's (2011) supervised approach) in the 0.107 range. Moreover, 63% (74 out of 117) of the segmented ROI images in the proposed method fall in the range $0 \leq error \leq 0.107$ compared to 32% (37), 37% (44), 49% (57) and 42% (49) for Kim's, Li and Ngan's (2007), Li and Ngan's (2011) and Chen and Li's methods, respectively. More significantly, only 5% (6 out of 117) of the segmented ROI images in the proposed method have an error greater than 0.25 compared to 44% (51), 31% (36), 28% (33) and 44% (51) for Kim's, Li and Ngan's (2007), Li and Ngan's (2011) and Chen and Li's methods, respectively. Also, the proposed method has a maximum error of 0.607 compared to 1 for the other four proposed methods.

TABLE I

PERFORMANCE EVALUATION BY OBJECTIVE CRITERION (SET 1)

Image	**Kim (2005) [41]**	**Li and Ngan (2007) [42]**	**Chen and Li (2012) [51]**	**Proposed Method**	**Max Error Difference (%)**
Dog	0.3435	0.0442	**0.012**	0.0156	0.36
Bee-Flower-1	0.4144	0.2366	**0.1087**	0.1188	1.01
Owl	0.1427	0.061	**0.0432**	0.0459	0.27
Strawberry	0.0598	0.0542	0.0356	**0.0345**	0.11
Beckham	0.0762	0.0279	0.0228	**0.021**	0.18
Flower-Yellow	0.1168	0.2262	0.075	**0.0194**	**5.56**
Average	0.1922	0.1084	0.0496	**0.0425**	

To provide an objective comparison of the three sets of images subjectively analysed earlier, the segmentation errors are computed and compared. Table I shows that for the first set (refer to Fig. 17) the proposed method performs better for 3 out of the 6 images while Chen and Li's

method performs better for the other 3 images. The average error across the 6 images for the proposed method is 0.043 compared to 0.05 for Chen and Li's method. The proposed method has a maximum improvement of 5.56% compared to 1.01% for Chen and Li's method.

TABLE II

PERFORMANCE EVALUATION BY OBJECTIVE CRITERION (SET 2)

Image	**Kim (2005) [41]**	**Li and Ngan (2007) [42]**	**Li and Ngan (2011) [52]**	**Proposed Method**	**Max Error Difference (%)**
Bee-Flower-2	0.2482	0.2173	**0.0674**	0.0997	3.23
Meerkat	0.3223	0.1368	0.1095	**0.0673**	**4.22**
Dragonfly	0.1563	0.1754	**0.0984**	0.0994	0.1
Flower-Red	0.1606	0.231	0.1047	**0.0795**	2.52
Frog	0.3145	0.0676	0.056	**0.0234**	3.26
Referee	0.1817	0.1035	0.0287	**0.015**	1.37
Average	0.2306	0.1553	0.0775	**0.0641**	

Table II shows that for the second set (refer to Fig. 18) the proposed method performs better for 4 out of the 6 images while Li and Ngan's (2011) method performs better for the other 2 images. The average error across the 6 images for the proposed method is 0.064 compared to 0.078 for Li and Ngan's (2011) method. Moreover, the proposed method has a maximum improvement of 4.22% compared to 3.23% for Li and Ngan's (2011) method.

TABLE III

PERFORMANCE EVALUATION BY OBJECTIVE CRITERION (SET 3)

Image	**Ye and Lu (2002) [54]**	**Kim (2005) [41]**	**Li and Ngan (2007) [42]**	**Liu et. al (2010) [53]**	**Proposed Method**	**Max Error Difference (%)**
Moth	0.16	0.0698	0.0589	0.085	**0.0263**	5.87
Cheetah	0.44	0.1828	0.1196	0.102	**0.0485**	5.35
Hornbill	0.12	0.133	0.1229	0.111	**0.0378**	**7.32**
Soccer	0.23	0.219	0.2937	0.177	**0.1677**	0.93
Average	**0.24**	**0.1512**	**0.1488**	**0.1188**	**0.0701**	

Table III shows that for the third set (refer to Fig. 19) the proposed method performs better for all 4 images compared to four of the other methods. Moreover, the average error across the 4

images is 0.07 with maximum improvement of 7.32%. The segmentation error is not one of the metrics used for an objective comparison in the fifth reference, Rafiee (2013) [55], and as a consequence those respective comparative results are not presented in Table III.

An objective comparison is made between the initial ROI estimation using the edge, gradient and HOS saliencies and the final ROI estimation that is based on a refinement of the former using *k*-means segmentation and CIEDE2000 colour difference analysis. From Table IV it is evident that the final method proposed for the extraction of ROI results in an optimal minimum error of 0.1058 – an improvement of 3.4% compared to the initial approximation of the ROI.

TABLE IV
PERFORMANCE EVALUATION BY OBJECTIVE CRITERION (ROI EXTRACTION)

Error	**Initial ROI**	**Final ROI**
0.036	19	24
0.107	48	50
0.179	21	24
0.250	12	13
0.321	8	2
0.393	3	1
0.464	1	0
0.536	2	1
0.607	1	2
0.678	1	0
0.750	0	0
0.821	0	0
0.893	0	0
0.964	0	0
1.000	1	0
Average	0.1398	0.1058

In addition to the objective comparison, the segmentation performance is computed by considering the average precision, recall and F-measure or balance F-score over the ground-truth databases. The general F-measure formula, denoted by F_β, is defined as

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall} \tag{6}$$

where

$$Precision = \frac{True\ Positive}{(True\ Postive + False\ Postive)} = \frac{|\Omega_{ROI} \bigcap \Omega_{GT}|}{|\Omega_{ROI}|} \tag{7}$$

and

$$Recall = \frac{True\ Positive}{(True\ Postive + False\ Negative)} = \frac{|\Omega_{ROI} \bigcap \Omega_{GT}|}{|\Omega_{GT}|}. \tag{8}$$

Achanta et. al [112] and Chen and Li [51] recommend a value of 0.3 for β^2. This is to allow for precision to be weighed more than recall.

TABLE V
AVERAGE F-MEASURE, PRECISION AND RECALL FOR 117 TEST IMAGES

Approach	**F-measure (%) for $\beta^2 = 0.3$**	**Precision (%)**	**Recall (%)**
Kim (2005) [41]	84.0	90.9	67.7
Li and Ngan (2007) [42]	84.6	88.1	75.3
Li and Ngan (2011) [52]	89.4	91.4	83.3
Chen and Li (2012) [51]	85.8	91.2	71.7
Rafiee (2013) [55]	91.3	**93.9**	83.6
Proposed Method	**95.2**	92.7	**96.1**

As evident in Table V, the proposed method achieves the highest average F-measure value of 95.2% as well as a recall (sensitivity) rate of 96.1%. The proposed method also exhibits a precision rate of 92.7%. Compared to Li and Ngan's (2011) method (nearest supervised rival), the proposed algorithm achieves gains of 5.8% and 1.3% in F-measure and precision, respectively. Compared to Rafiee's (2013) method (nearest unsupervised rival), the proposed algorithm achieves gains of 3.9% and 12.5% in F-measure and recall, respectively, and exhibits a 1.2% difference in precision.

Owing to space constraints only a selected number of the total ROI extraction results are presented in this section. The author may be contacted at serenr@gmail.com for more than 300 LDOF images and their associated segmentation results; over 230 of the images also include their associated binary ground-truth masks.

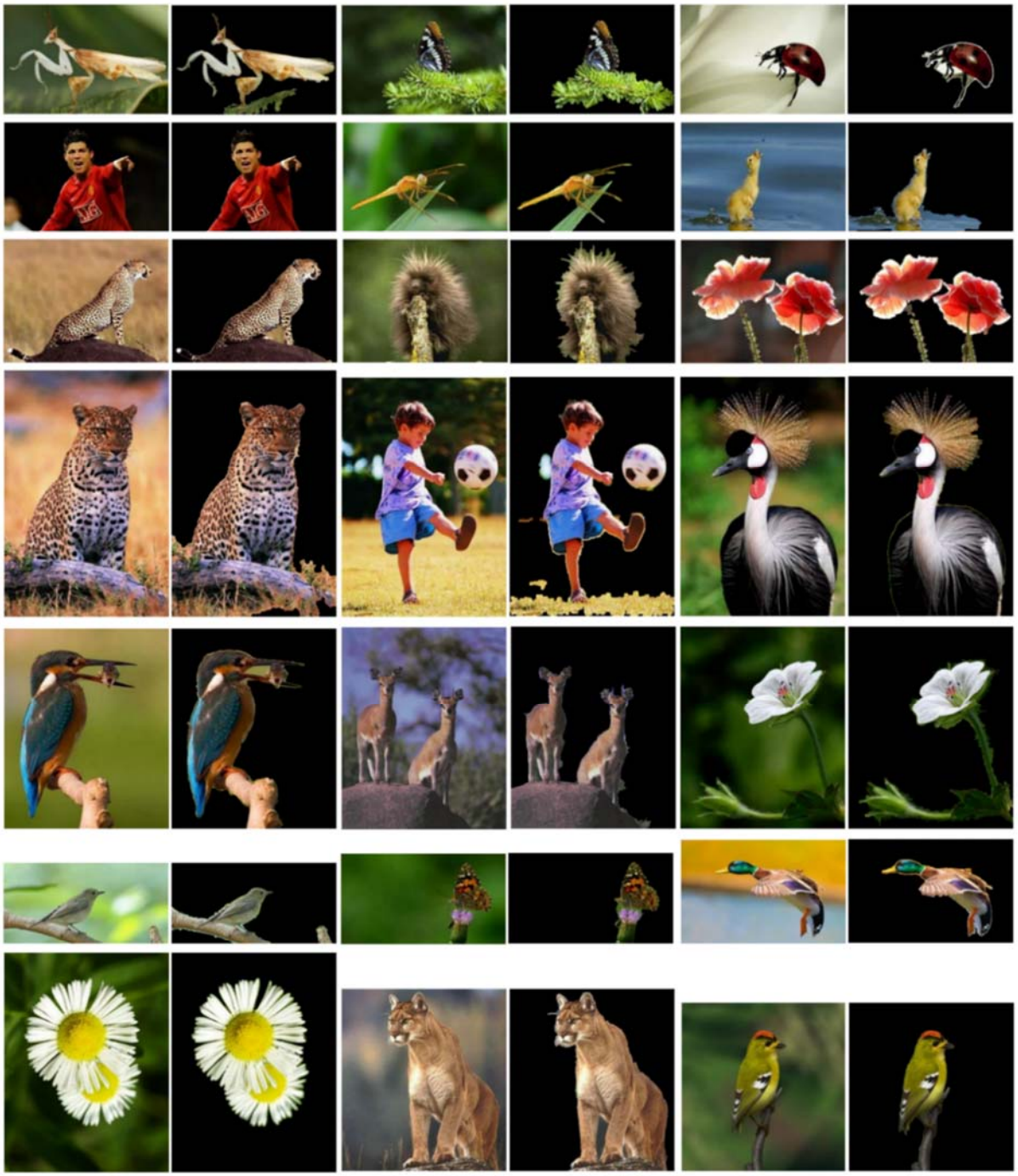


Fig. 21 Selection of segmentation results (Set 1).

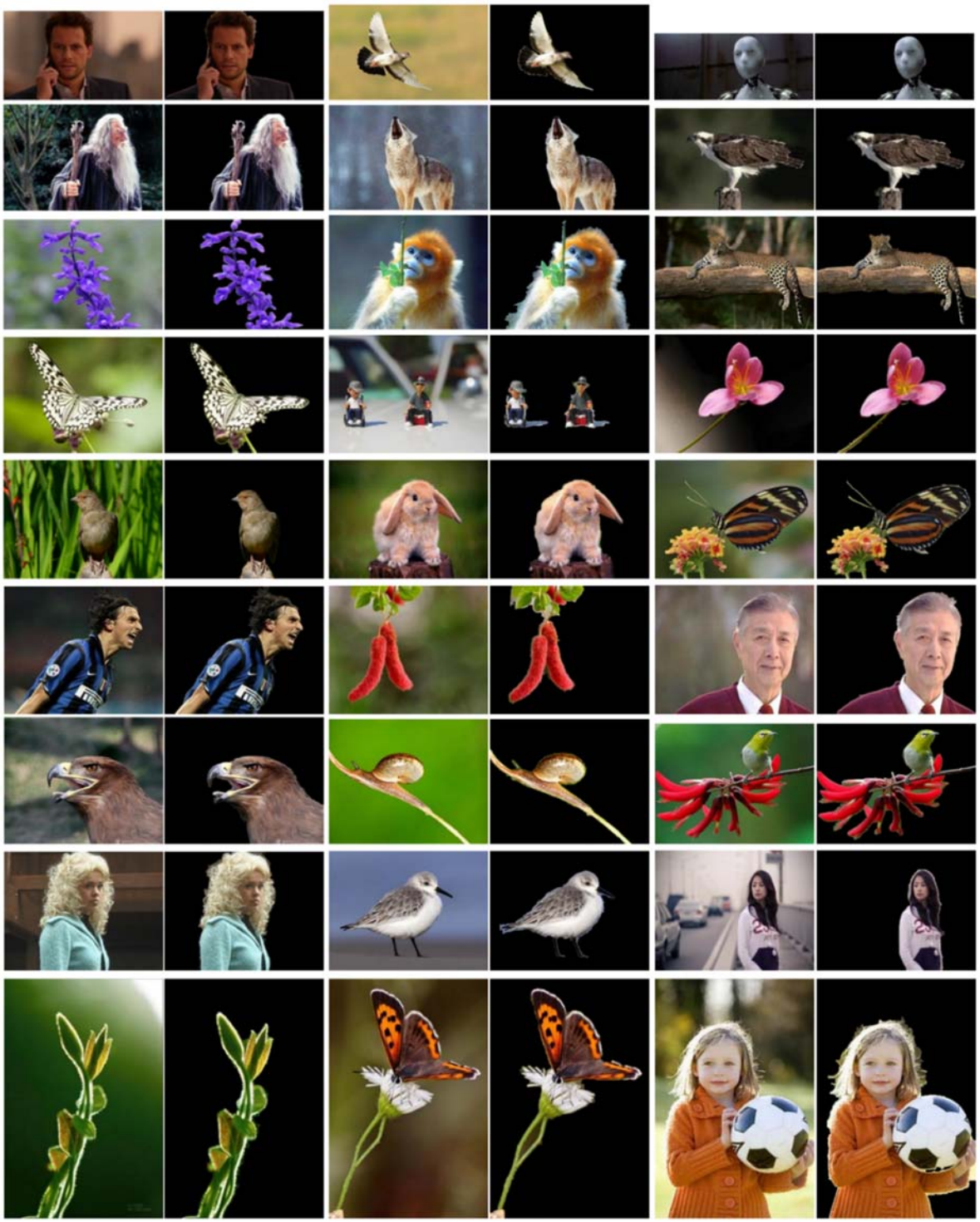


Fig. 22 Selection of segmentation results (Set 2).

For a truer reflection of the proposed algorithm for unsupervised ROI extraction some of the limitations (failed or inadequate segmentation results) are considered. From the 117 test images, 6 of them produced segmentation errors greater than 0.25. These are illustrated in Fig. 23.

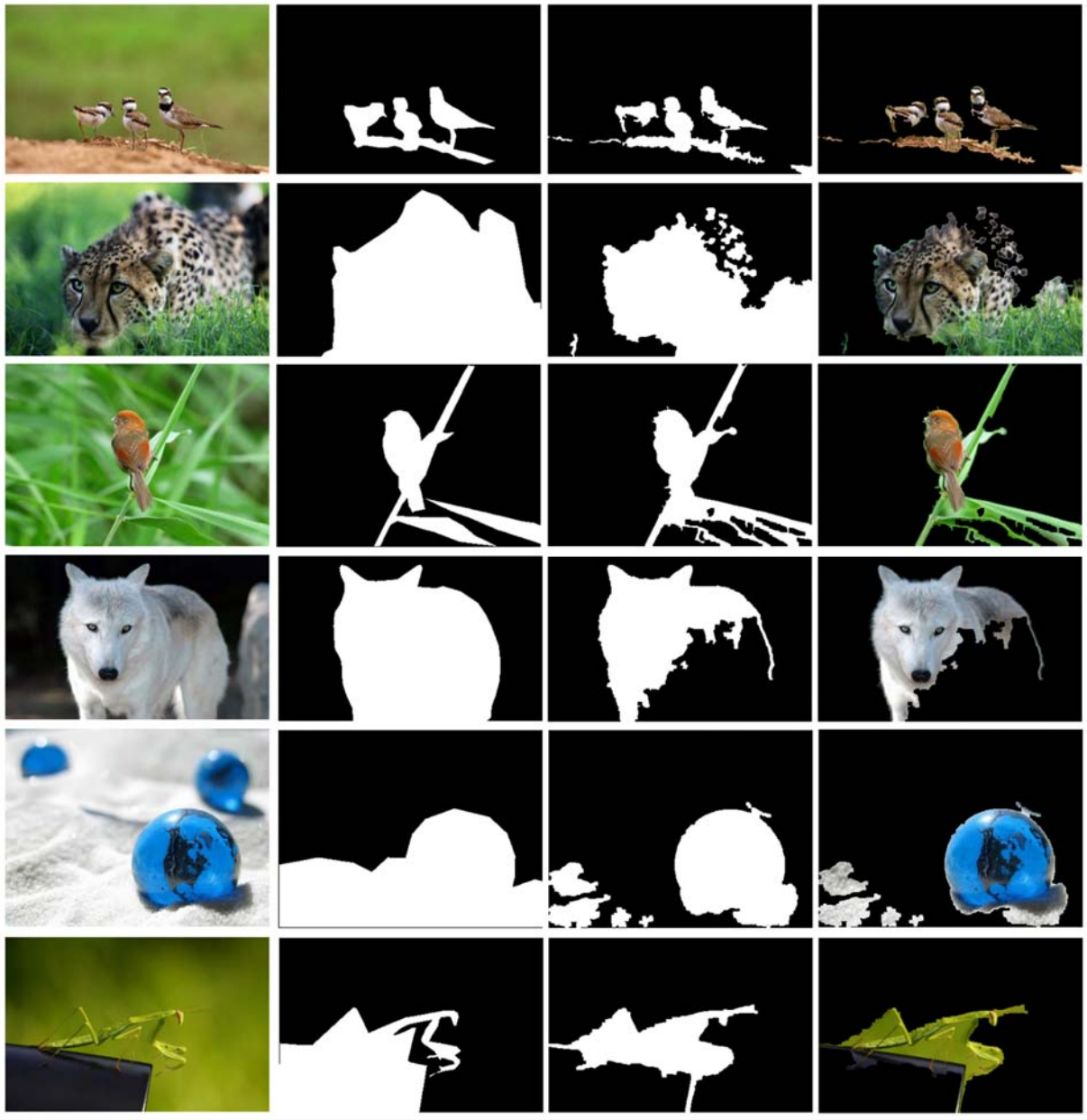


Fig. 23 Segmentation results from the 117 test images with error greater than 0.25: *Three-Birds*, *Crouching-Cheetah*, *Bird-Reed*, *Albino-Wolf*, *Marbles* and *Mantis* (from top to bottom).

The LDOF technique in which the cheetah is photographed in the *Crouching-Cheetah* image results in the head being more in focus than the rear of the body. As a consequence the proposed model incorrectly classifies the rear of the cheetah as belonging to the BG. Furthermore, the out

of focus leaves in front of the lower jaw result in that region exhibiting low frequency saliency values. As a consequence the lower jaw is also misinterpreted as belonging to the BG. The leaf located below the bird in the *Bird-Reed* image is wrongly classified as the BG. Even though the OOI (the bird) is effectively reflected and the result is subjectively acceptable, the large false negative value skews the segmentation error. The praying mantis in the *Mantis* image is perched on top of a post box in the ROI. However, the glossy paint as well as the curvature of the roof of the post box results in that region exhibiting low frequency saliency values. As a consequence the region is misclassified as belonging to the BG and, owing to the associated large false negative value, unevenly skews the segmentation error.

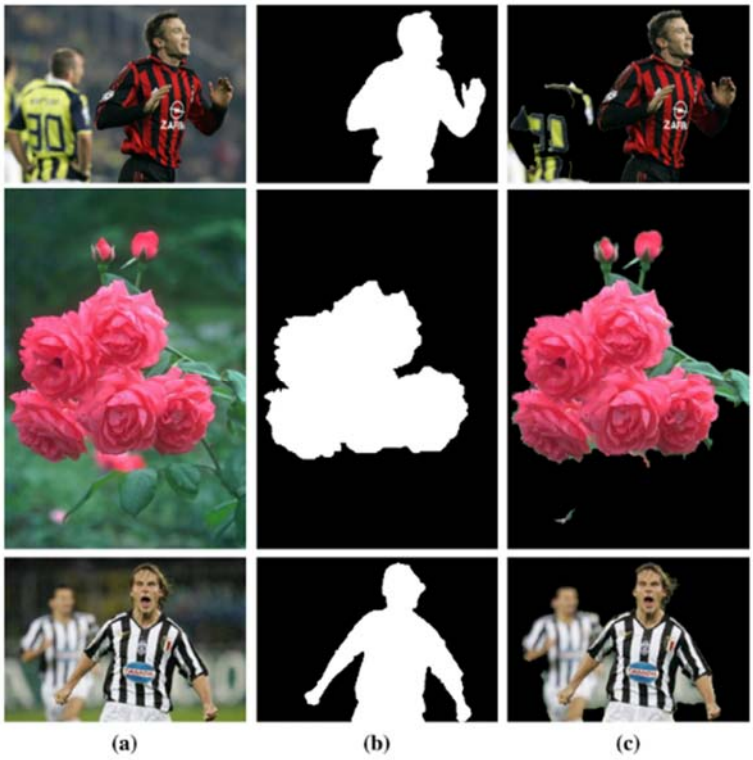


Fig. 24 Segmentation results for images with low defocus of foreground versus background [51]. (a) Original images; (b) Associated ground-truth masks; (c) Results of proposed algorithm.

Fig. 24 illustrates some segmentation failures resulting from images with low differences in the defocus degree of the foreground versus background. Referring to the image on the first row, it may be observed that although the OOI (soccer player in red and black jersey) is accurately segmented the soccer player in the BG with the yellow and black jersey is misclassified as belonging to the FG. This results in a large false positive segmentation error. A possible resolution to this error may be to consider retaining only the largest segment. With the image in the second row the error may be subjectively explained. Although the ground-truth excludes both the unopened buds above, as well as the leaves to left of the bunch, both these regions may be

subjectively considered as belonging to the ROI. In the case of the image in the third row, there is a similar colour distribution between the OOI (the soccer player in the FG with his arms spread apart) and the soccer player running in the semi-blurry BG region. The segmentation component responsible for both the expansion as well as the refinement of the ROI misinterprets both these similar colour regions as belonging to the FG.

Some of the large segmentation errors may be a consequence of incorrectly delineated ground-truths. However, most of the limitations highlighted above reflect some of the challenges still inadequately resolved in the arena of unsupervised ROI extraction.

The proposed algorithm is computationally benchmarked against four of the other proposed models. The testing computer is a Quad Core CPU 2.50 GHz with 4.00 GB of RAM. The proposed method is implemented using MATLAB R2014a.

TABLE VI
AVERAGE COMPUTATIONAL SPEED
FOR 117 TEST IMAGES (AVG. SIZE 400×300)

Approach	**Average Computational Speed (in seconds)**
Kim (2005) [41]	8
Li and Ngan (2007) [42]	40
Li and Ngan (2011) [52] – supervised approach	4.7
Chen and Li (2012) [51]	7
Rafiee (2013) [55]	2.3
Proposed Method	4.8

Table VI provides a comparative analysis of the average computational cost across the 121 test images. In this regard, although the method proposed by Rafiee (2013) is shown to be more computationally efficient, the proposed unsupervised model achieves an average processing time of 4.8 s, which is comparable to Li and Ngan's (2011) supervised method (4.7 s on average). In addition, the proposed algorithm provides at least a 42% improvement on Kim's as well as Chen and Li's methods (8 s and 7 s, respectively) and a significant improvement on Li and Ngan's (2007) method (approximately 40 s on average).

Even though the dataset used for the validation may appear to be somewhat limited, the robustness of the proposed model may possibly be further strengthened by testing using images coming from different cameras with various resolution sizes. This may be considered at a later stage.

Moreover, a concern with the HOS technique is that although it may be effective in exploiting the high frequency attributes in LDOF images it may often exhibit sensitivity to noisy images. A method to address these concerns has been proposed by Ahn and Chong (2015) [113] whereby adaptive second-order statistics are considered. For improved robustness this adaptive approach may possibly be incorporated at a later stage into the proposed model.

E. Summary

The proposed model has been both subjectively as well as objectively measured against several state-of-the-art methods [41, 42, 51-55] through the comparison of 121 LDOF test images selected from two common datasets.

By subjectively examining Fig. 17, Fig. 18 and Fig. 20, the results shows that although there exists some minor false positives, the proposed approach is nevertheless comparable to the other seven approaches. From the subjective examination of Fig. 17 the results show that the proposed model is comparable to Chen and Li's (2012) method while providing a noticeable improvement to Kim's (2005) and Li and Ngan's (2007) methods. Moreover, by subjectively examining Fig. 18, the results again show that the proposed technique is comparable to Li and Ngan's (2007) unsupervised method as well as Li and Ngan's (2011) supervised method while demonstrating a noticeable improvement to Kim's (2005) method. Furthermore, the subjective examination of Fig. 19 provides evidence of the proposed model being comparable to the approaches proposed by Kim (2005), Liu et. al (2010) and Rafiee (2013) while providing a noticeable improvement to Ye and Lu's (2002) as well as Li and Ngan's (2007) methods.

As is objectively evident from Fig. 20, the results show that the proposed method outperforms the compared ROI extraction methods. Although the proposed unsupervised method shows 4% (24 compared to 25) and 25% (24 compared to 32) less accuracy in the 0.036 range against Chen and Li's (2012) unsupervised method and Li and Ngan's (2011) supervised method, respectively, it nevertheless outperforms all other methods by at least twice the accuracy in the 0.107 range. Moreover, 63% (74 out of 117) of the segmented ROI images in the proposed unsupervised method fall in the range $0 \leq error \leq 0.107$ compared to 32% (37), 37% (44) and 42% (49) for Kim's (2005) [41], Li and Ngan's (2007) [42] and Chen and Li's (2012) [51] unsupervised approaches, respectively, as well as 49% (57) for Li and Ngan's (2011) [52] supervised method. More significantly, only 5% (6 out of 117) of the segmented ROI images in the proposed unsupervised approach have an error greater than 0.25 compared to 44% (51), 31% (36) and 44% (51) for Kim's (2005) [41], Li and Ngan's (2007) [42] and Chen and Li's (2012) [51] unsupervised methods, respectively, as well as 28% (33) for Li and Ngan's (2011) [52] supervised approach. Also, the results show that the proposed method has a maximum error of 0.607 compared to 1 for the other four proposed methods.

As is objectively evident from Table V, the proposed method achieves both the highest average F-measure value of 95.2% and recall rate of 96.11%. In addition a precision rate of 92.7% is achieved. Compared to best unsupervised approach (Chen and Li (2012) [51]) gains of 9.4% and 1.5% are attained for the average F-measure and precision rate, respectively. Compared to best supervised approach (Rafiee (2013) [55]) a gain of 3.9% is attained for the average F-measure.

As evident from Table VI, the computational time for the proposed algorithm is comparable to the supervised approach of Li and Ngan (2011) [52] and achieves at least a 43% reduction compared to the other unsupervised approaches of Chen and Li (2012) [51], Kim (2005) [41] and Li and Ngan (2007) [42].

F. Conclusion

Effective unsupervised image segmentation remains a challenging problem in the field of image and video processing. The difficulty with unsupervised region segmentation is that the problem is ill-posed. This is owing to having no primary segmentation criteria. This research presents a novel method for the unsupervised extraction of the ROI from a single LDOF image.

The basic premise of this research involves the interrogation of high-frequency components for use in ROI extraction. However, a concern is that by only exploiting the high-frequency components inconsistencies may occur when interpreting both the defocused and focused elements of an image. This is owing to some areas of the defocussed regions exhibiting busy texture attributes with high enough frequency components so as to be considered as in focus, and some areas of focussed regions containing constant texture characteristics with insufficiently high-frequency components so as to be considered as out of focus.

The proposed method reduces the effects of these inconsistencies by considering a correlated approach to the ROI detection and extraction problem through the interrogation of several saliencies based on gradient and higher-order statistics (HOS) analysis techniques. In addition to the correlation of these saliencies the proposed model supplements these approaches through edge and segmentation correlation.

The proposed method consists of four stages. The first stage involves the extraction of edge, gradient and HOS saliency attributes. The second stage involves the extrapolation of an ROI reference mask based on the gradients in the image. The third stage involves the interrogation of the gradient and HOS saliencies under the guidance of the ROI reference mask and the subsequent extraction of the ROI. The fourth and final stage involves the refinement of the ROI. These refinements are performed through a closer inspection of the outer regions using *k*-means segmentation as well as the border regions using CIEDE2000 colour difference analysis.

The proposed algorithm has been assessed both subjectively as well as objectively and is shown to be comparable or outperform existing unsupervised state-of-the-art ROI extraction

approaches [41, 42, 51-55]. Results show that compared to the best unsupervised approach a gain of 3.9% is attained for the average F-measure. Although the proposed method is built on existing techniques, it nevertheless extends as well as introduces several novel concepts into the ROI extraction framework.

Future work may include investigating further strategies for saliency estimation and correlation as well as the expansion of the model to incorporation 2D LDOF sequential video and real-time functionality.

III. UNSUPERVISED MATTING OF THE OBJECT-OF-INTEREST IN LOW DEPTH-OF-FIELD IMAGES

The explicit extraction of the region-of-interest (ROI) from within a low depth-of-field (LDOF) image is discussed in the previous section on *p*. 18. This component of the research expands on the ROI extraction concept for specific scenarios of LDOF images and considers the more refined extraction of the object of interest (OOI) from within the ROI.

A. Introduction

The closed boundary delineation of the OOI is often vital in arenas such as machine vision, segmentation, rotoscoping and 2D-to-3D conversion. For all LDOF images there will exist regions in focus (focused) and regions out of focus (defocused). This is illustrated in Fig. 25. The focused regions, which are typically referred to as the foreground (FG) or the so-called ROI, are associated with high-frequency (sharp) components and the defocused regions, which are typically referred to as the background (BG), are association with low frequency (blurred) components. For most LDOF images the key purpose is to give more prominence to an object, such as a person or animal, the so-called OOI, rather than a region within a scene.

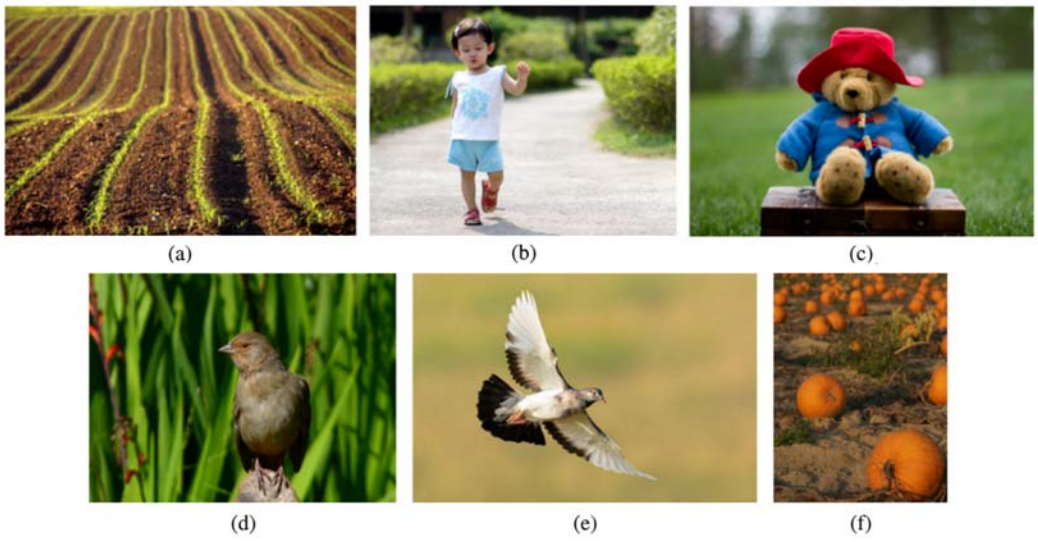


Fig. 25 Low depth-of-field images [60, 63]

Some of the variations commonly encountered with LDOF images are illustrated in Fig. 25. These include several of the permutations that may exist for the ROI and ROI-OOI combinations. For the ROI there are two very broad LDOF image classifications. The first is where no distinct OOI is present within the ROI, as in the case of a landscape, prairie or field, like for example the image illustrated in Fig. 25 (a). The second is where there exists a distinct OOI within

the ROI, like for example the images illustrated in Fig. 25(b)–(f). Within this second classification two further ROI-OOI permutations or categories may exist. The first category is where the entire ROI represents the OOI, as illustrated in Fig. 25(d) and (e) as well as Fig. 26(a). The second category is where the ROI contains both an OOI as well as a distinct ground region, as illustrated in Fig. 25(b), (c) and (f).

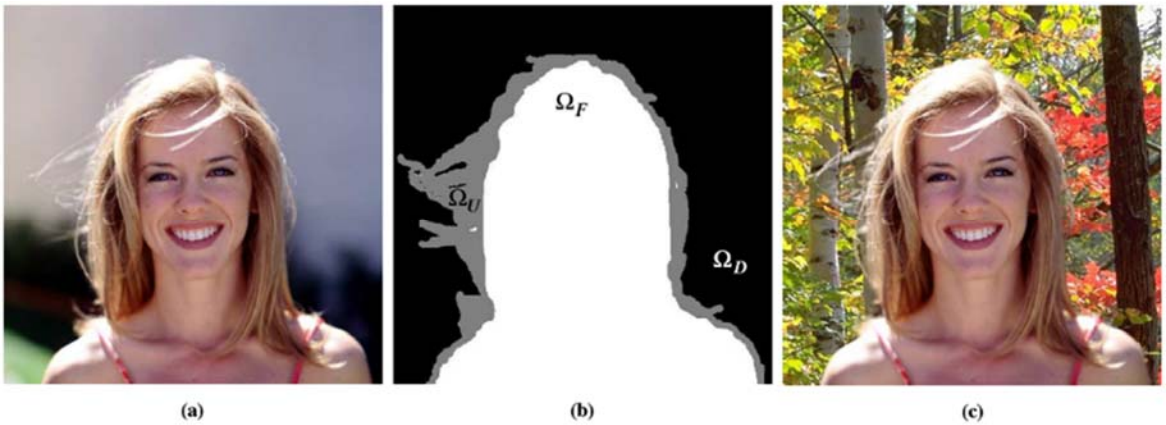


Fig. 26 Image Matting [114]. (a) Original defocused image; (b) Trimap; (c) Matted foreground with new background.

The term image matting or just *matting* is used to generically describe the soft extraction or isolation of an in focus OOI from the surrounding out of focus BG [114]. This is illustrated in Fig. 26. Matting more often than not refers explicitly to the scenario where the OOI constitutes the entire ROI i.e. the first ROI-OOI category. However, the ROI may be more generically described as representing a region rather than a distinct closed boundary object within a LDOF image and the OOI, if it exists, may be defined as being an element of the ROI and either constituting the entire ROI or denoting a subcomponent of the ROI. The problem with this latter scenario (the second ROI-OOI permutation) is that since the in focus OOI is contained within the in focus ROI, interrogating only the high-frequency components may be inadequate in distinguishing between the OOI and the non-OOI region within the ROI.

There are three approaches to image matting. These include manual, semi-automatic and unsupervised. For the second approach, some intermediate user interaction is usually required for the identification, delineation and refinement of the OOI. This is often in the form of a predefined trimap [45]. For the third unsupervised matting approach it may be necessary to autonomously generate a sort of conceptual trimap within the model.

A trimap is a tri-colour (black, white and grey) image indicating definite background regions, definite foreground regions and unclassified pixels, respectively, with the unclassified regions most likely located around the borders of significant objects. This is illustrated in Fig. 26(b) where Ω_F refers to the *definitely focused* OOI, Ω_D denotes the *definitely defocused* BG region

and Ω_U refers to the *unknown* region. This latter region is primarily associated with the boundary regions of the OOI and represents a selection of unclassified pixels like, for example, thin wisps of hair or fur that may be interpreted either as belonging to the foreground or the background. These complex pixels would require additional examination to provide a more precise alpha matte. For the second ROI-OOI permutation the unknown region Ω_U in the trimap needs to account for both the border regions of the OOI as well as the non-OOI region.

For an unsupervised approach, owing to the ill-posed nature of the problem, there is no a priori reference for the focussed OOI and the surrounding focussed non-OOI or ground-plane region, denoted by GOI. As a consequence in this research the trimap is not specifically employed as a reference for the delineation of the OOI from the surrounding image, but instead it may be more appropriately described as a tri-region classification of the LDOF image. In this case Ω_F represents a more generic description of the in focus OOI, Ω_U provides a broad description of the in focus ground plane region that either partially or entirely surrounds the OOI, and Ω_D represents a more general description of the out of focus BG region.

This research expands on the matting concept by proposing a method for the automatic delineation of the closed boundaries of an OOI from a LDOF image where the OOI represents a subcomponent of a larger ROI. This will require firstly, the extraction of the ROI and secondly, a reference for the OOI based on an analysis of the gradients in the image and thirdly, the estimation, refinement and subsequent delineation of the OOI from the rest of the ROI through the interrogation and correlation of edges and gradients in the ROI together with cluster-based segmentation.

Only images where part of the OOI overlaps with the out of focus BG of the image are considered in this research. Scenarios where the OOI is completely enveloped by the ROI, like for example the image illustrated in Fig. 25(f), are excluded. In these latter scenarios, absent of any a priori information pertaining to the OOI or other monoscopic cues such as motion, there is no discernible robust means of differentiating the OOI from the non-OOI region in the ROI.

This paper is organised as follows: Section B provides a description of the proposed model as well as the necessary steps required to perform unsupervised OOI matting. Section C reports the experimental results. Section D provides a brief discussion on some of the limitations of the research and section E closes the paper.

B. Proposed Approach

The proposed model involves four stages. The first stage considers the extrapolation of the in focus FG or the region-of-interest (ROI) from the out of focus BG. This stage is described in detail in the previous section *Unsupervised Region-of-Interest Extraction based on the Correlation of Gradient and Higher-Order Statistics Saliencies* on *p*. 18. The second stage

involves the extrapolation of a baseline reference for the OOI. The third stage involves the generation of a more precise estimate of the OOI by expanding the OOI baseline reference using edges and *k*-means segmentation. For each approximation of the OOI, starting with the baseline reference, the LDOF image is classified into one of six proposed ROI-OOI permutations or types. The fourth and final stage involves the precise delineation of the close boundaries of the OOI from rest of the ROI.

1) OOI Baseline Reference

In an unsupervised scenario no a priori baseline reference exists for either the OOI or GOI and as a result the problem becomes ill-posed. To resolve this conundrum, the proposed model uses the dominant gradient saliencies in the image to initially derive a baseline reference for the OOI. The objects in an image are expected to exhibit more clearly defined or dominant edges than non-object regions. Moreover, for scenarios where a distinct OOI exists, the boundary of this region is expected to contain more clearly defined or dominant edges than the rest of non-OOI region (including the GOI) within the LDOF image.

In this research, dominant edges are considered as a subcategory of dominant regions. These regions are high energy connected components that are represented by edges with the longest lengths as well as those forming closed segments. Since each of these dominant regions represent individual connected components containing a certain amount of energy (pixels), it may be assumed with a high degree of probability that majority of the closely clustered high energy regions within the ROI are associated with the OOI.

In the proposed model the gradient saliency map (Π_{GRAD}) defined on *p*. 26 is considered as the reference for the dominant regions. The OOI baseline reference, denoted by Π_{OOI}^{INIT}, is determined by firstly, considering only connected components within Π_{GRAD} that intersect with the ROI and secondly, excluding any segment within this range that is below a certain energy threshold. From experiment, any segment or filled region having a pixel count (energy) less than 5% of the perimeter of the image is omitted. This is illustrated in Fig. 27(c).

2) LDOF Image Classification

In the proposed model OOI matting is only performed on LDOF images where the OOI represents a subcomponent of a larger ROI. As a consequence it is necessary at this stage to unambiguously distinguish between the various types of LDOF image scenarios. Although it is implausible to account for every possible scenario, a wide range of LDOF images may nevertheless be classified by considering only a few of the most common variations encountered, such as those illustrated in Fig. 25

For most scenarios involving an OOI the focus of the image is centred on the OOI. However, in some cases the OOI may be located on top of a ground region and depending on the focal plane settings during image capture the in focus ROI may also encapsulate (a portion of) this ground region. This is denoted by GOI. Each LDOF image classification may be described as a permutation of three regions of concern. These include the in focus OOI, in focus GOI and out of focus BG regions.

In the proposed model, if a LDOF image contains a distinct OOI, then the location of the OOI is also considered as a variable. This results in six possible types of LDOF image classifications. At this stage the purpose of the classification is not to yield a precise delineation of the OOI but rather to provide a technique of broadly classifying the LDOF image into one of the six possible scenario types. Table VII provides a description of these various classification types together with their trimap representations. The grey region in the trimap represents an approximation of the GOI, the white region denotes an approximation of the OOI and the black region represents an approximation of the BG (non-ROI).

Based on the descriptions provided in Table VII this research only explores OOI matting for images classified as either type 2 or 3. Type 1 and 6 either contains no OOI or the OOI is indistinguishable from the ROI and is therefore not applicable. For type 4 and 5 classifications the method proposed for ROI extraction in the previous section on *p*. 18 may be used for OOI extraction without requiring any further adaptation.

The terms OOI and GOI refer to the *precise* delineation of both the OOI and the ground region within the ROI. However, for the purpose of LDOF image classification it is only necessary to have a rudimentary approximation of these two sub-regions. The type of the LDOF image classification is determined by interrogating the OOI and the ground plane approximations and associating the permutation with a corresponding description provided in Table VII.

TABLE VII

DESCRIPTION OF VARIOUS TYPES OF LDOF IMAGES

Classification	**Description**	**Trimap**
TYPE 1	▪ The ROI extends across the entire width of the image ▪ A distinct OOI is NOT present ▪ The GOI represents the entire ROI ▪ This is illustrated in Fig. 25(a).	
TYPE 2	▪ The FG extends across the entire width of the image ▪ A distinct OOI is present ▪ A distinct GOI is present ▪ The OOI overlaps with the BG region ▪ No part of the OOI overlaps with the bottom row of the image ▪ This is illustrated in Fig. 25(d). The ground plane may either be less than the width or span the entire width of the image.	
TYPE 3	▪ The FG extends across the entire width of the image ▪ A distinct OOI is present ▪ A distinct GOI is present ▪ The OOI overlaps with the BG region ▪ The OOI overlaps with the bottom row of the image ▪ This is illustrated in Fig. 25(e). The ground plane may either be less than the width or span the entire width of the image.	

Classification	Description	Trimap
TYPE 4	▪ A distinct OOI is present ▪ A distinct GOI is NOT present ▪ The OOI represents the entire ROI ▪ The OOI overlaps with the bottom row of the image ▪ This is illustrated in Fig. 25(c).	
TYPE 5	▪ A distinct OOI is present ▪ A distinct GOI is NOT present ▪ The OOI represents the entire ROI ▪ This is illustrated in Fig. 25(b).	
TYPE 6	▪ The ROI extends across the entire width of the image ▪ A distinct OOI is present ▪ A distinct GOI is present ▪ The OOI is entirely enveloped by the ROI ▪ This is illustrated in Fig. 25(f). When the OOI is fully enveloped by the GOI the location of the OOI within the GOI is not considered as a variable.	

At this point a rudimentary classification of the image may be made. The OOI is chosen as the rectangular bounded region of the OOI baseline reference Π_{OOI}^{INIT} (Refer to Fig. 27(c)). For the ground plane region, the rectangular bounded region of Π_{OOI}^{INIT} is initially omitted from the original ROI (Refer to Fig. 27(b)) and any segment not incident with the bottom border of the image is subsequently excluded; this is denoted by Π_{GRND}^{INIT}.

From experiment, if any part of Π_{GRND}^{INIT} is incident on either the left or right boundaries of the image, then the image is rudimentarily classified as either type 2, 3 or 6. If Π_{OOI}^{INIT} is fully enveloped by Π_{GRND}^{INIT} , then it is a type 6 classification. If none of the previous conditions apply and if the bottom border of Π_{OOI}^{INIT} is located on the bottom border of the image, then it is a type 3 classification, else it is a type 2 classification.

If Π_{OOI}^{INIT} spans the entire width of the image or $\Pi_{OOI}^{INIT} = \varnothing$, then there is no discernible a priori means of determining whether the FG represents a landscape type situation with no distinct OOI or if the entire ROI is the OOI. In this eventuality the image is classified as type 1.

If $\Pi_{GRND}^{INIT} = \varnothing$, then this implies a type 4 or 5 classification. However, if the bottom border of Π_{GRND}^{INIT} is located on the bottom border of the image, then it is a type 4 classification, else it is a type 5 classification.

A classification of the LDOF image is made each time an approximation of the OOI is produced. If the image is classified as either type 1, 4, 5 or 6, then no further processing occurs and the OOI is equated to the ROI.

For illustrative purposes, a classification tri-map may be constructed. Initially, if $\Pi_{GRND}^{INIT} \neq \varnothing$, then by using the centre column of the image, Π_{GRND}^{INIT} is split into two subcomponents. Subsequently a straight line bounded representation of Π_{GRND}^{INIT} is constructed by protracting a straight line through the centroids of the upper most row of both subcomponents across the entire span of the image.

For certain LDOF image scenarios the ground plane may often appear skewed (non-horizontal) and inconsistent or incomplete. This is owing primarily to the camera parameters and/or the position of the camera during image capture. Although the focus of the OOI is minimally affected in these scenarios, a consequence is that some of the defocussed regions of the ground plane may be mistakenly interpreted as belonging to the BG thereby resulting in an inconsistent ROI region. As a consequence, if the protracted straight line intersects the bottom border of the image or if only one subcomponent exists, the bounded ground plane is generated by protracting the upper most row of Π_{GRND}^{INIT} across the entire width of the image.

Finally, the classification tri-map is produced by setting the pixels associated with the bounded region of Π_{GRND}^{INIT} to 128 and those associated with rectangular bounded region of Π_{ROI}^{OOI} to 255. The remaining pixels, which represent an approximation of the BG region, are all assigned the value 0. This is illustrated in Fig. 27(d).

Fig. 27 Image Matting. (a) Original defocused image; (b) Region-of-Interest; (c) Baseline OOI reference; (d) LDOF image classification.

3) OOI Matting

The baseline OOI reference Π_{OOI}^{INIT} may be used for a rudimentary classification of the LDOF image. However, on its own, it is inadequate for the precise delineation of the OOI from the ROI. For this purpose a more refined estimate of the OOI and GOI is required. In the proposed model the matting of the OOI is achieved over four stages. The first two stages involve the estimation of the OOI which is derived by starting with Π_{OOI}^{INIT} and subsequently expanding on this region through a series of incremental assessments. The third stage involves an estimation of the ground plane region by using the estimation of the OOI as a reference. The fourth and final stage involves the matting of the OOI by excluding the ground plane region GOI from the ROI.

Fig. 28 Image Matting. (a) Original LDOF image; (b) Initial OOI reference; (c) Final OOI reference.

The first stage initially uses Π_{OOI}^{INIT} as a foundation and subsequently expands this region using common edges. In the proposed model the edge saliency map (Π_{EDGE}) defined on *p*. 26 is considered for this purpose. Initially, any segment in Π_{EDGE} having a certain overlap within a buffer range of Π_{OOI}^{INIT} is considered as belonging to the OOI. From experiment the rectangular bounded region of Π_{OOI}^{INIT} is selected as the buffer range and a percentage overlap of 30% is used as the threshold. Subsequently, close neighbourhood edges are linked and any closed boundary regions are filled. This expanded OOI region is denoted by $\Pi_{OOI'}^{1}$.

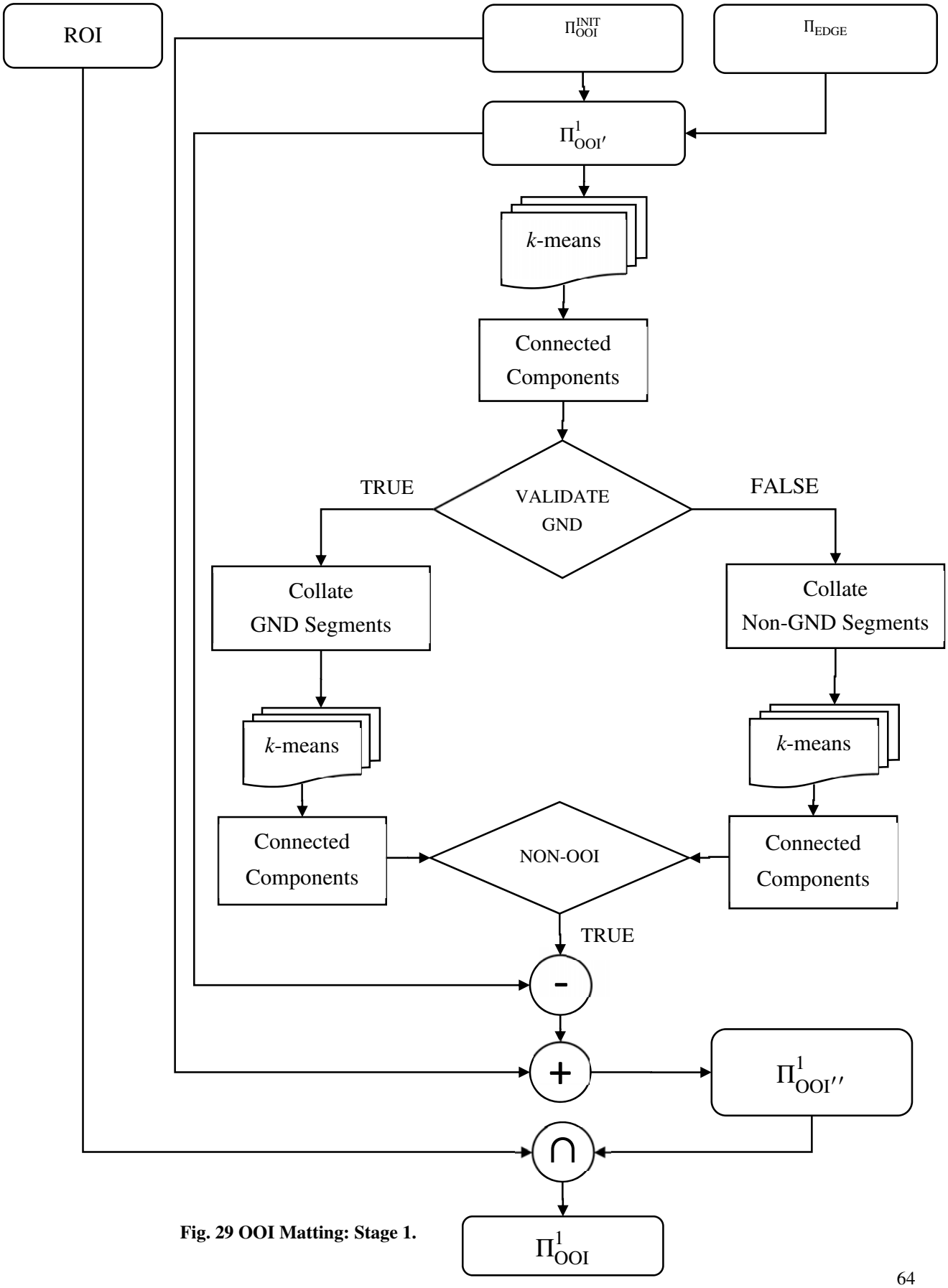


Fig. 29 OOI Matting: Stage 1.

The next step involves a deeper interrogation and refinement of the new regions in $\Pi^1_{OOI'}$. Firstly, $\Pi^1_{OOI'}$ is segmented using *k*-means clustering. The *k*-means segmentation technique employed in this research is discussed on *p*. 35. The output is only dependent on the a priori number of clusters requested. Owing to the nature of LDOF images some segmentation discontinuities may arise when objects and regions transition between the FG and BG regions. From experiment, a cluster value of 3 is shown to minimise under-segmentation while also avoiding over-segmentation. For the various OOI matting substages a cluster value of 3 is set whenever *k*-means segmentation is employed.

Secondly, each connected component in each of the cluster arrays are analysed, separated and collated in arrays containing ground regions and non-ground regions. From experiment, by employing a type of Gestalt reasoning [24], any connected component that is incident with the bottom border of the image is initially considered as belonging to the ground region.

Thirdly, the two collated arrays are further sub-segmented using *k*-means clustering. Subsequently the connected components within each of the sub-segmented cluster arrays are validated. From experiment any connected component having a percentage overlap of less than 85% of the rectangular bounded region of Π^{INIT}_{OOI} is considered as invalid and excluded from $\Pi^1_{OOI'}$. To ensure the integrity of the OOI baseline reference is preserved a union is performed between Π^{INIT}_{OOI} and the refined $\Pi^1_{OOI'}$; the result is denoted by $\Pi^1_{OOI''}$.

Finally, the initial OOI reference Π^1_{OOI} is derived by intersecting $\Pi^1_{OOI''}$ with the ROI. A flow diagram of the algorithm for the first stage is provided in Fig. 29 and an illustration of the results are provided in Fig. 28(b).

The second stage involves the expansion of Π^1_{OOI}. Firstly, the entire LDOF image is segmented using *k*-means clustering. Secondly, the connected components within each of the segmented cluster arrays are validated. From experiment any connected component is considered as belonging to the OOI if either it has a percentage overlap of greater than 85% of the rectangular bounded region of Π^1_{OOI} or if more than 50% of the perimeter of the connected component is adjacent to Π^1_{OOI}. Thirdly, the invalid regions are collated into a single array and subsequently segmented using *k*-means clustering. Again the connected components within each of the clusters are assessed using the aforementioned validation criteria with the expanded OOI region as the reference. The last process is repeated two more times for the invalid segments and at each validation step the expanded OOI region is used as the reference. This expanded OOI region is denoted by $\Pi^2_{OOI'}$. A flow diagram of the algorithm for the first component of the second stage is provided in Fig. 30.

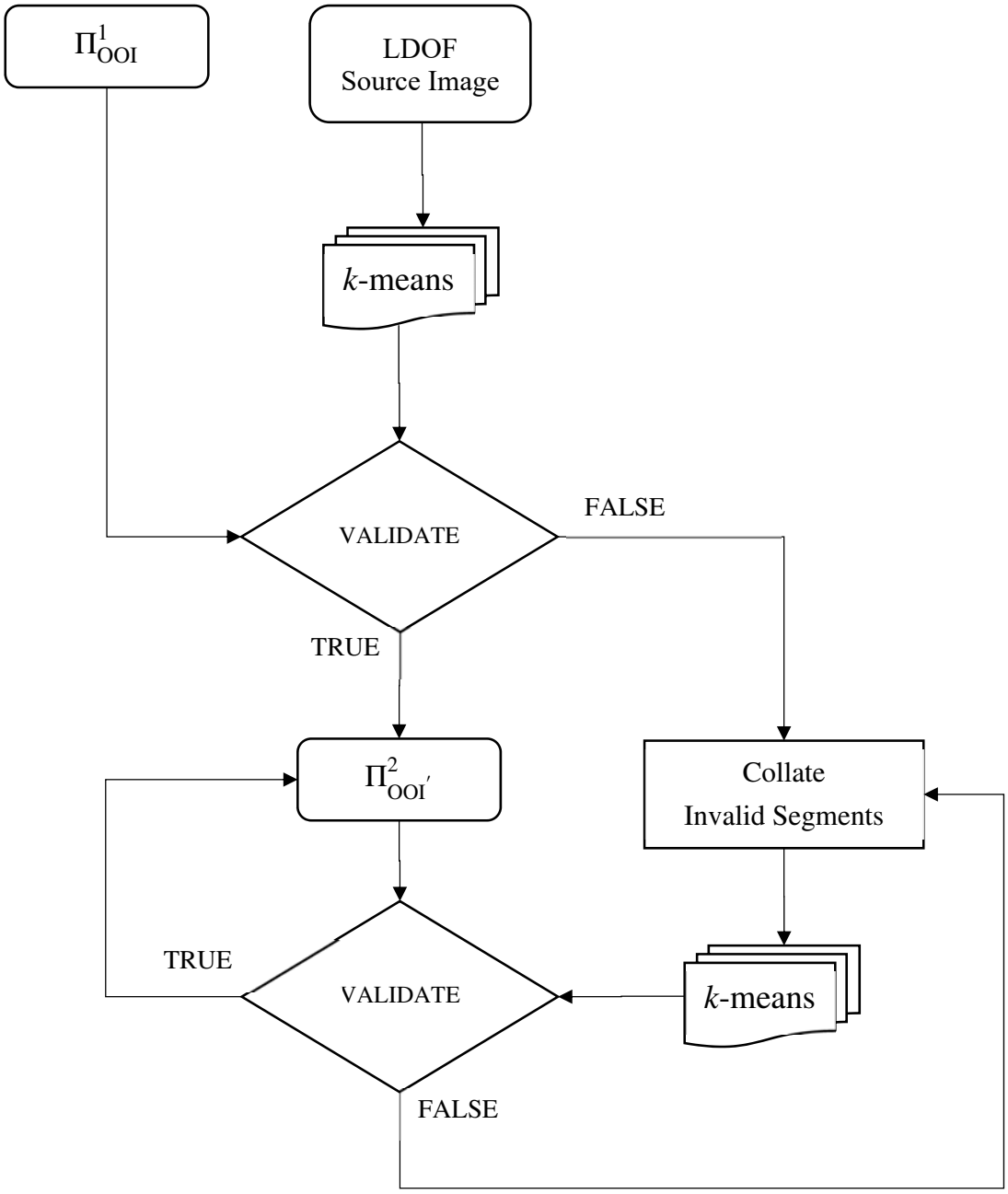


Fig. 30 OOI Matting: Stage 2a.

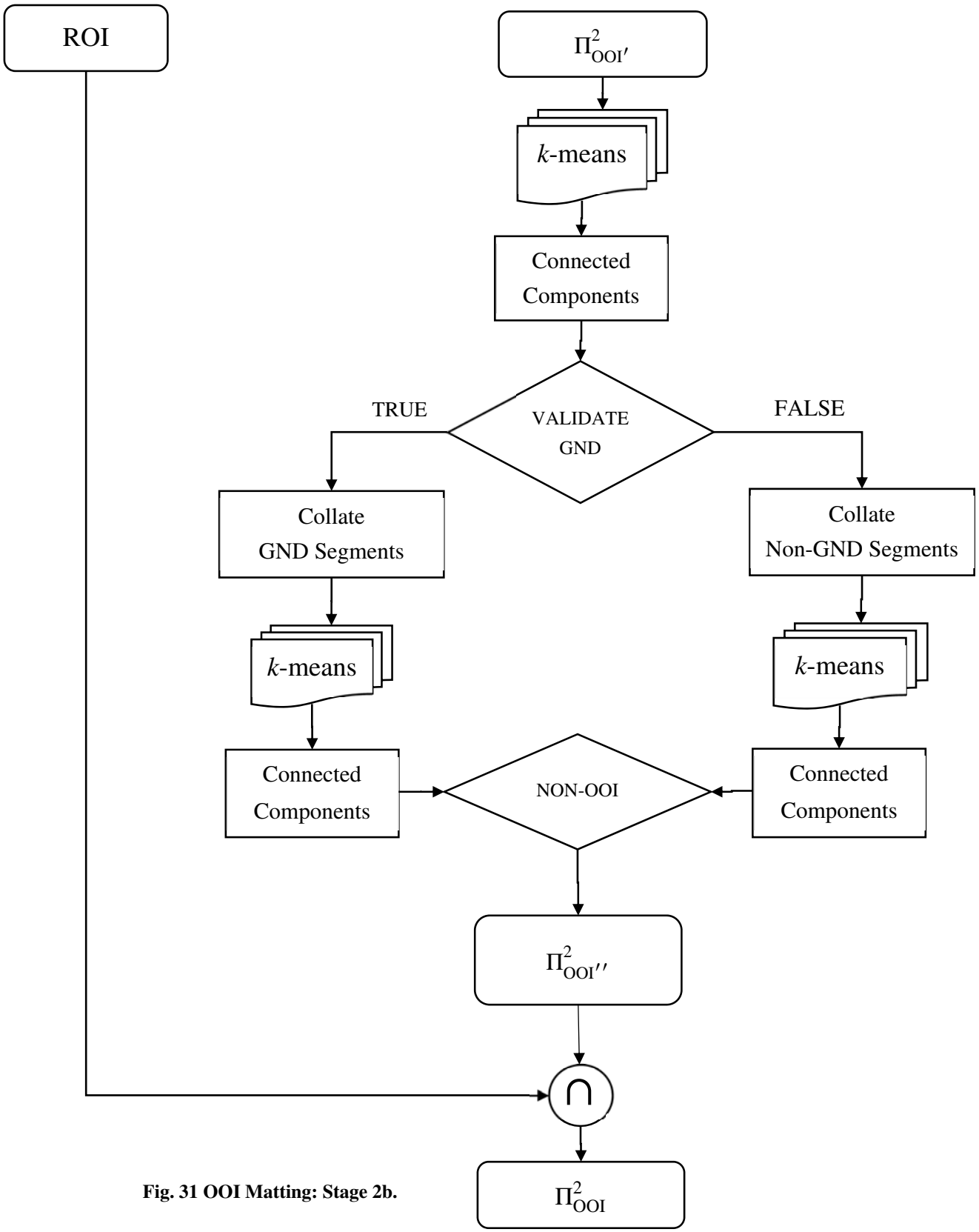


Fig. 31 OOI Matting: Stage 2b.

The next step involves a deeper interrogation and refinement of the new regions in $\Pi^2_{OOI'}$ similar in manner to the first stage. Firstly, $\Pi^2_{OOI'}$ is segmented using *k*-means clustering. Secondly, each connected component in each of the cluster arrays are analysed, separated and collated in arrays containing ground regions and non-ground regions. Ground regions include any connected components that are incident with the bottom border of the image. Thirdly, the two collated arrays are further sub-segmented using *k*-means clustering. Subsequently the connected components within each of the sub-segmented cluster arrays are validated. From experiment any connected component having a percentage overlap of less than 85% of the rectangular bounded region of Π^1_{OOI} is excluded from $\Pi^2_{OOI'}$; this is denoted by $\Pi^2_{OOI''}$.

Finally, the refined OOI region Π^2_{OOI} is derived by performing an intersection between $\Pi^2_{OOI''}$ and the ROI. A flow diagram of the algorithm for the second component of the second stage is provided in Fig. 31 and an illustration of the results are provided in Fig. 28(c).

Fig. 32 Image Matting. (a) Original defocused image; (b) ROI; (c) Ground region GOI; (d) Matted OOI.

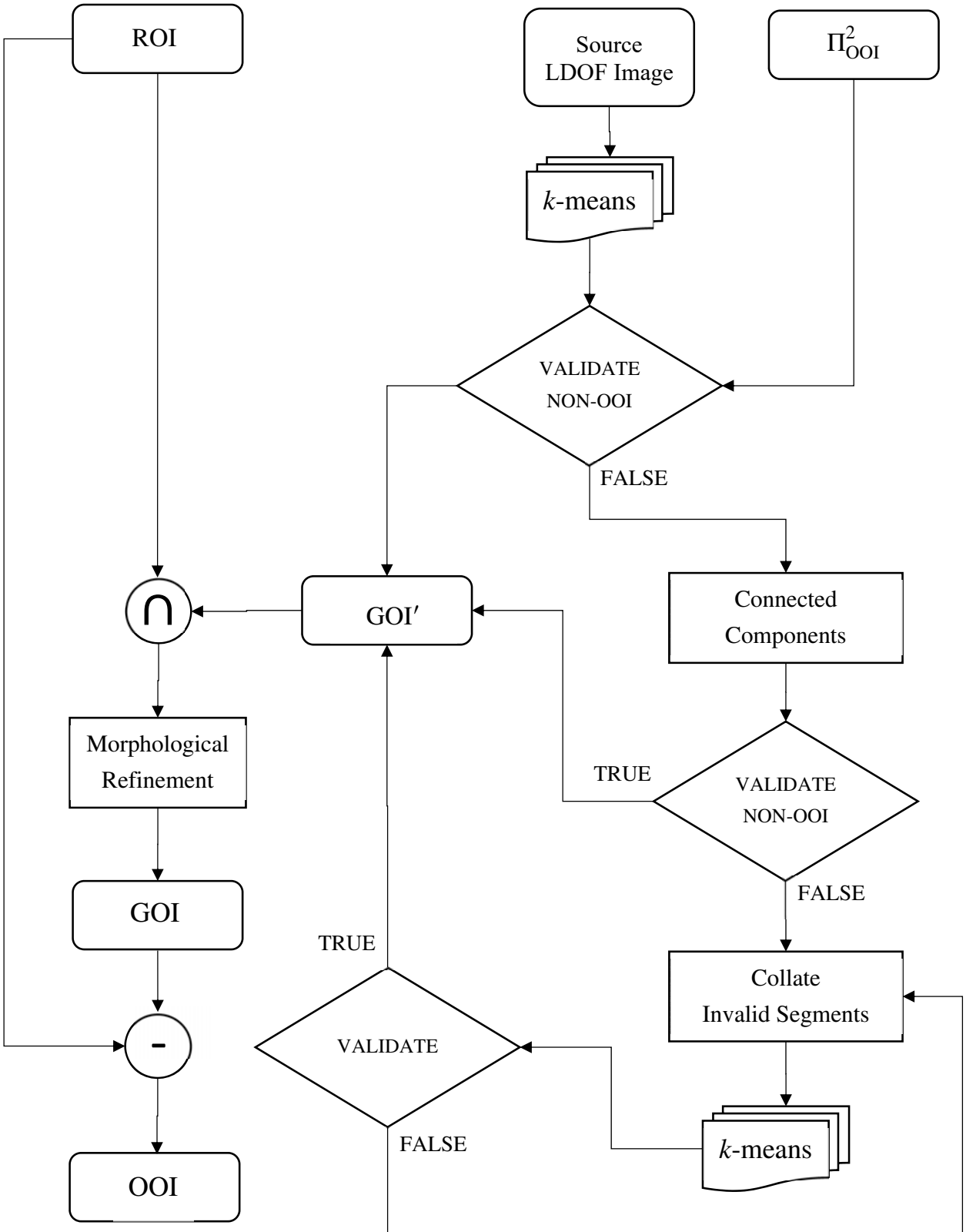


Fig. 33 OOI Matting: Stages 3 and 4.

The third stage involves the extraction of the ground region by estimating the non-OOI region of the entire LDOF image. From experiment, the following method is employed for this purpose. Firstly, the entire LDOF image is segmented using *k*-means clustering. Secondly, an entire cluster array is considered to belong to the non-OOI region if it has less than a 5% overlap with the with the OOI reference Π^2_{OOI}. Thirdly, if a cluster arrays is invalid, then the individual connected components within the cluster array are assessed. In this case any connected component having less than a 10% overlap with the Π^2_{OOI} is also considered as belonging to the non-OOI region. Fourthly, any cluster array or connected component excluded from the non-OOI region are subsequently collated and reassessed. This is achieved by performing segmentation of the collated non-OOI region using *k*-means clustering. The first two stages above are subsequently applied to these sub-segmented cluster arrays. To account for the reproducible inconsistencies associated with *k*-means segmentation the entire process above is iterated 3 times with the non-OOI region being constructed in an accumulative manner. Finally, any segment in the non-OOI region not located on the border region of Π^2_{OOI} or completely enveloped by Π^2_{OOI} is excluded.

The final in focus ground plane region GOI is extrapolated by performing an intersection between the non-OOI region and the ROI. Any low energy regions such as sparse edges and small connected components are considered to be invalid and are excluded from GOI. From experiment the morphological erosion and subsequent dilation of the region using a 5×5 square structuring element is adequate for the refinement of these regions. Moreover, any of the remaining connected components that are not incident with the bottom border of the image are excluded from GOI. Subsequently, the matted OOI is derived by excluding GOI from the ROI. A flow diagram of the combined algorithm for the third and fourth stages is provided in Fig. 33. and an illustration of the results of the GOI and the matted OOI are provided in Fig. 32.

C. Results

There are three distinct outcomes to this research. The first involves the rudimentary approximation of the OOI by using the dominant gradients in the image as a baseline reference. The second involves the use of the rudimentary approximation of the OOI together with the extracted ROI to classify the LDOF image into one of six possible types. The third involves the matting of the OOI from the ROI for type 2 and 3 classifications.

The proposed methods are tested on real world images. These include images from Li and Ngan (2007) [42] (together with their ground-truth masks), the COREL dataset [110, 111] as well as numerous images from the World Wide Web (WWW). A separate set of training images were

also sourced from Li and Ngan (2007) [42], the COREL dataset [110, 111] , Bae and Durand [60], Ko et. al [65], Zhou and Sim [63] and from WWW.

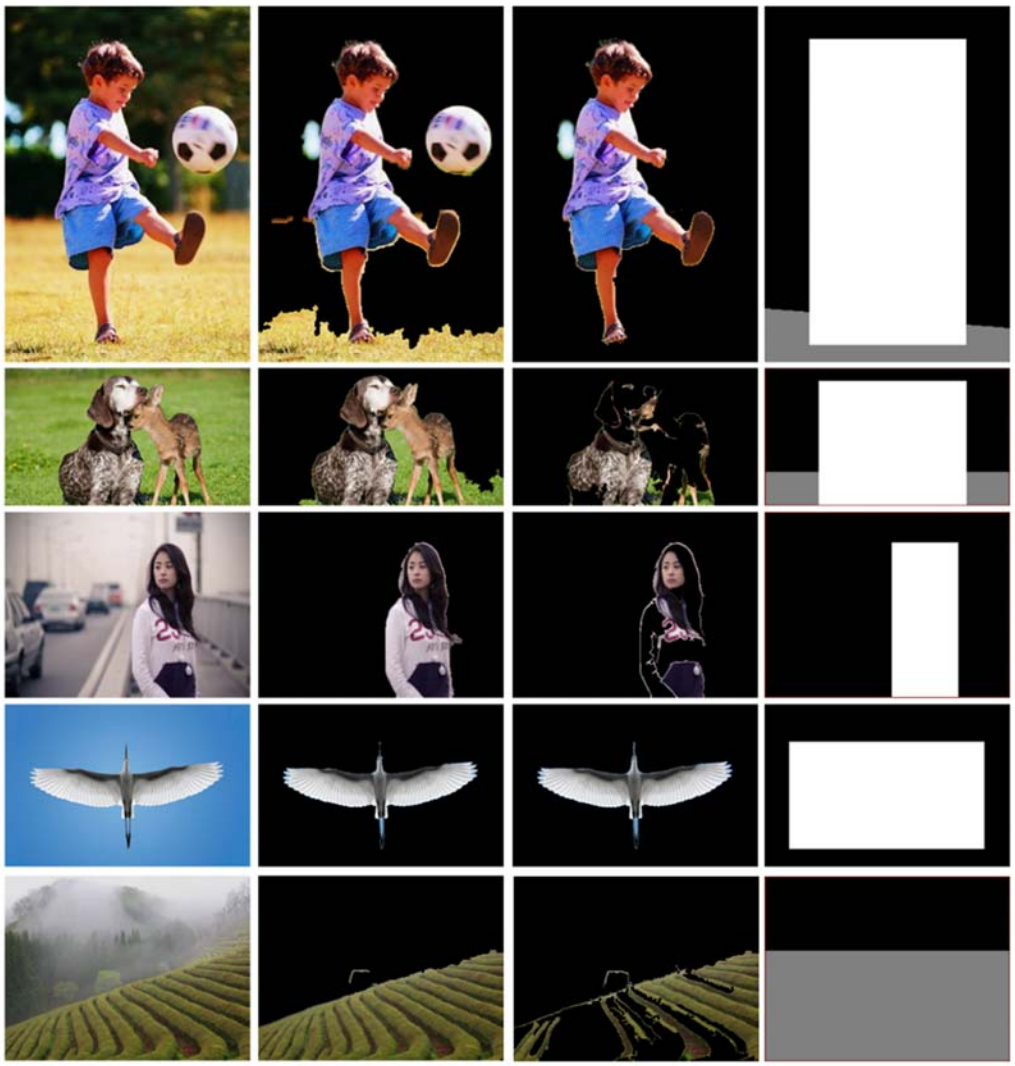


Fig. 34 Experimental results for LDOF image classification: *Boy-Soccer*, *Dog-Deer*, *Girl-Bridge*, *Crane* and *Field* (from top to bottom). (a) Original defocused image; (b) ROI; (c) OOI reference; (d) Classification.

Some of the results obtained for the methods proposed for the rudimentary approximation of the OOI and LDOF image classification are shown in Fig. 34. The original images in Fig. 34(a) are titled (from top to bottom) as *Boy-Soccer*, *Dog-Deer*, *Girl-Bridge*, *Crane* and *Field*, respectively. Fig. 34(b) provides the corresponding extrapolated ROIs. These are derived using the method proposed in the previous section *Unsupervised Region-of-Interest Extraction based on the Correlation of Gradient and Higher-Order Statistics Saliencies* on *p*. 18. Fig. 34(c) shows the corresponding rudimentary approximation of the OOI proposed for use in the classification of the LDOF image. Fig. 34(d) provides the results obtained using the proposed method for LDOF

image classification. The OOI is indicated in white, the ground region in grey and the (defocussed) background in black. The descriptions as well as templates of the six proposed types of classifications are provided in table VII on *p*. 59.

Although the approximations of the OOI for the *Boy-Soccer* and *Crane* images may be considered as reasonably accurate closed boundary delineations of the OOI, the *Dog-Deer*, *Girl-Bridge* and *Field* images are noticeably inadequate and further expansion and correlation is required in order to have consistency across a wide range of LDOF images. Nevertheless, the results show that a reasonable enough approximation of the OOI is derivable from the dominant gradients in the image and that this is sufficient for the accurate classification of a LDOF image.

Some of the results obtained for the method proposed for OOI matting of type 2 and 3 classified LDOF images are shown in Fig. 35. The original images in Fig. 35(a) are titled (from top to bottom) as *Stag*, *Model*, *Puppies*, *Baseball-Girl* and *Tiger-Water*, respectively. Fig. 35(b) provides the corresponding extrapolated ROIs. These are derived using the method proposed in the previous section *Unsupervised Region-of-Interest Extraction based on the Correlation of Gradient and Higher-Order Statistics Saliencies* on *p*. 18. Fig. 35(c) shows the matted ground region of the ROI. Fig. 35(d) provides the results obtained using the proposed method for OOI matting of single 2D LDOF images.

Although the girl in the background of the *Baseball-Girl* image may be considered as an error, this is actually a consequence of the ROI sub-stage not effectively dealing with low differences in the defocus degree of the foreground versus background and not as a result of the proposed OOI matting method. This also applies to the small region between the horns of the stag in the *Stag* image. Similarly, the tiger in the *Tiger-Water* image should exclusively be considered as the OOI. However, the water region is significantly contrasted from the surrounding ground region and, absent of any a priori knowledge, should also be considered as belonging to the OOI.

The results show that by correlating the ROI with the dominant gradients in the image together with cluster-based segmentation it is possible to produce a reasonably accurate delineation of the closed boundary region of the OOI. Although the proposed method is inherently dependent the accuracy of the segmentation, the use of the dominant gradients provides a baseline reference for the OOI and as a consequence deal with the ill-posed problem associated with the unsupervised correlation of these segments.

Objective analysis of category types 4 and 5 LDOF images are presented in the previous section *Unsupervised Region-of-Interest Extraction based on the Correlation of Gradient and Higher-Order Statistics Saliencies* on *p*. 37. Currently in the literature there is no objective OOI/non-OOI ground-truth data available for category types 2 and 3 LDOF images. As a consequence the presented results for these specific category types are only subjectively assessed. Future work may include the generation of an OOI/non-OOI ground-truth database for all

categories of LDOF images. Although the generation itself will be subjectively produced it may be objectively assessed through peer consensus.

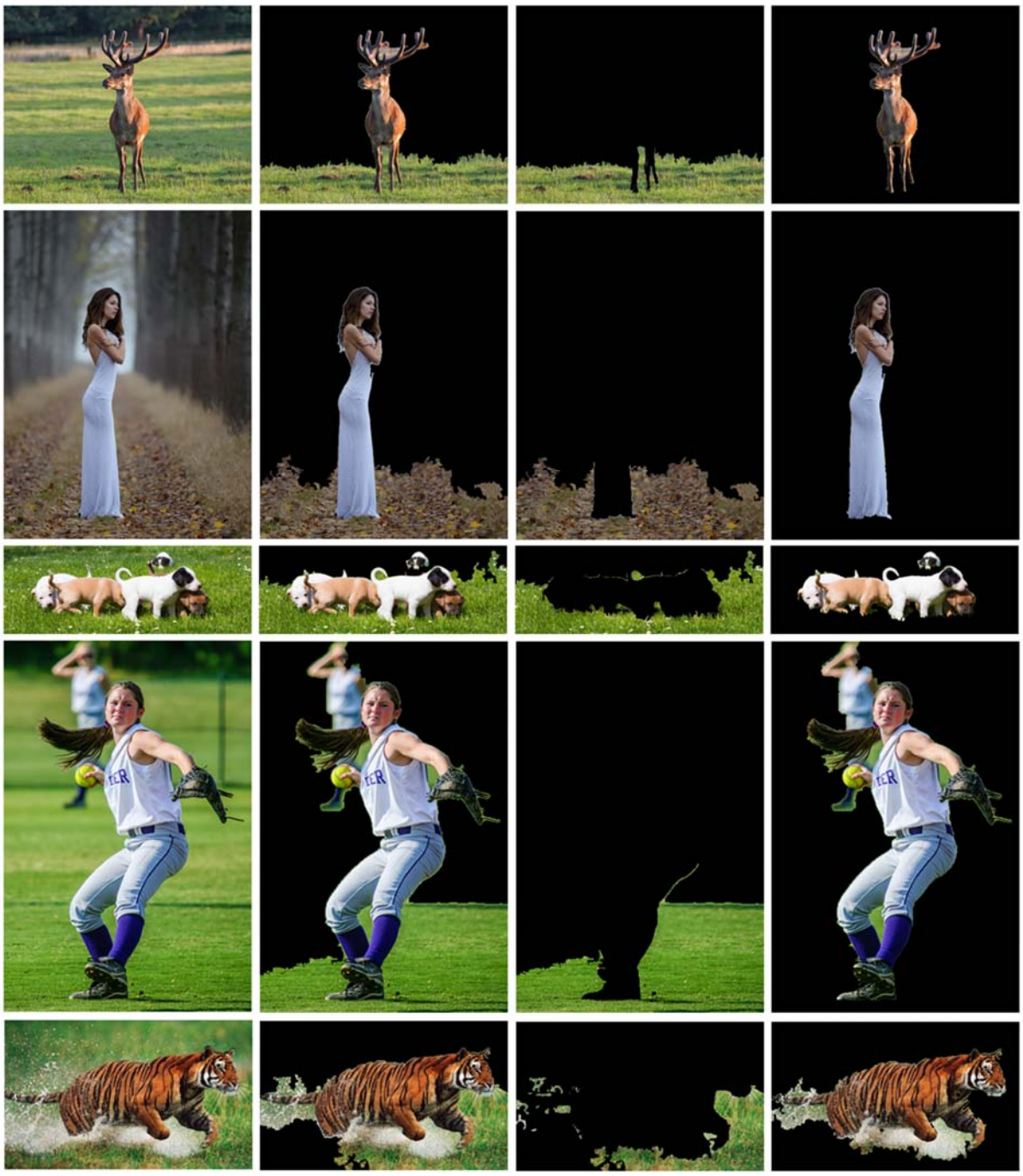


Fig. 35 Image matting results 1: *Stag*, *Model*, *Puppies*, *Baseball-Girl* and *Tiger-Water* (from top to bottom). (a) Original defocused image; (b) ROI; (c) Matted ground region; (d) Matted OOI.

Owing to space constraints only a selected number of the total OOI matting results are illustrated in this section. The author may be contacted at serenr@gmail.com for more than 300 LDOF images and their associated segmentation results; over 230 of the images also include their associated binary ground-truth masks.

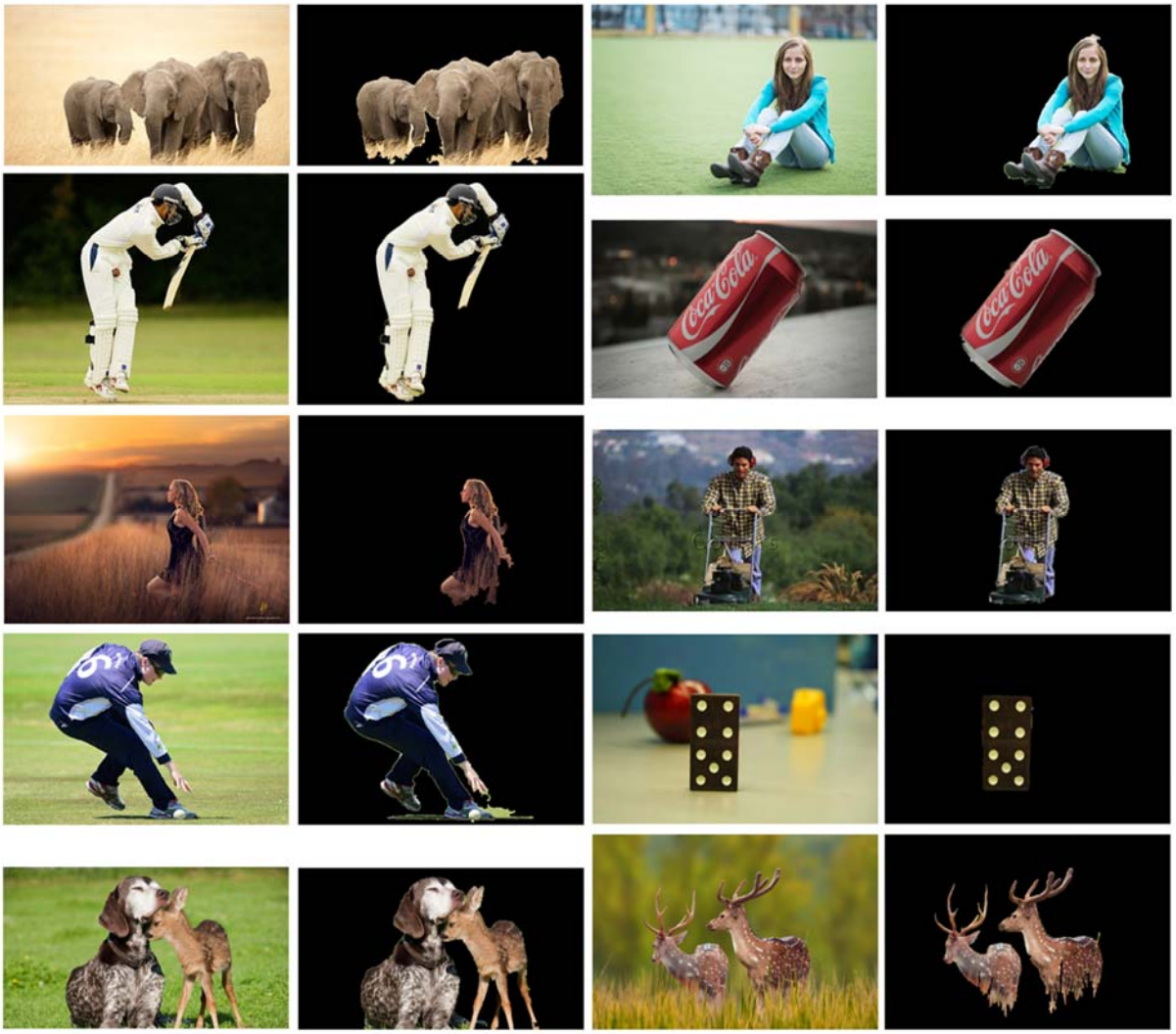


Fig. 36 Image matting results 2.

D. Discussion

A notable drawback of the proposed method is the that accuracy of the matting is dependent on the accuracy of delineated segmentation clusters. For some scenarios where the OOI may exhibit similar intensity values to the ground region the segmentation may result in some regions of the OOI being mistaken as belonging to the surrounding ground region and vice versa. These are illustrated in the *Girl-Tracks* and *Girl-Hitchhiker* images in Fig. 37.

Although the delineation of the ROIs as well as the classification of the images are accurate, the matting results are shown to be in error. In the *Girl-Tracks* image, owing to the colour of the ground and the boots being relatively close in intensity, the *k*-means clustering algorithm associates both regions with the same cluster. In the *Girl-Hitchhiker* image, the shadow cast from the OOI creates regions of continuity with the ROI and discontinuity with the ground plane.

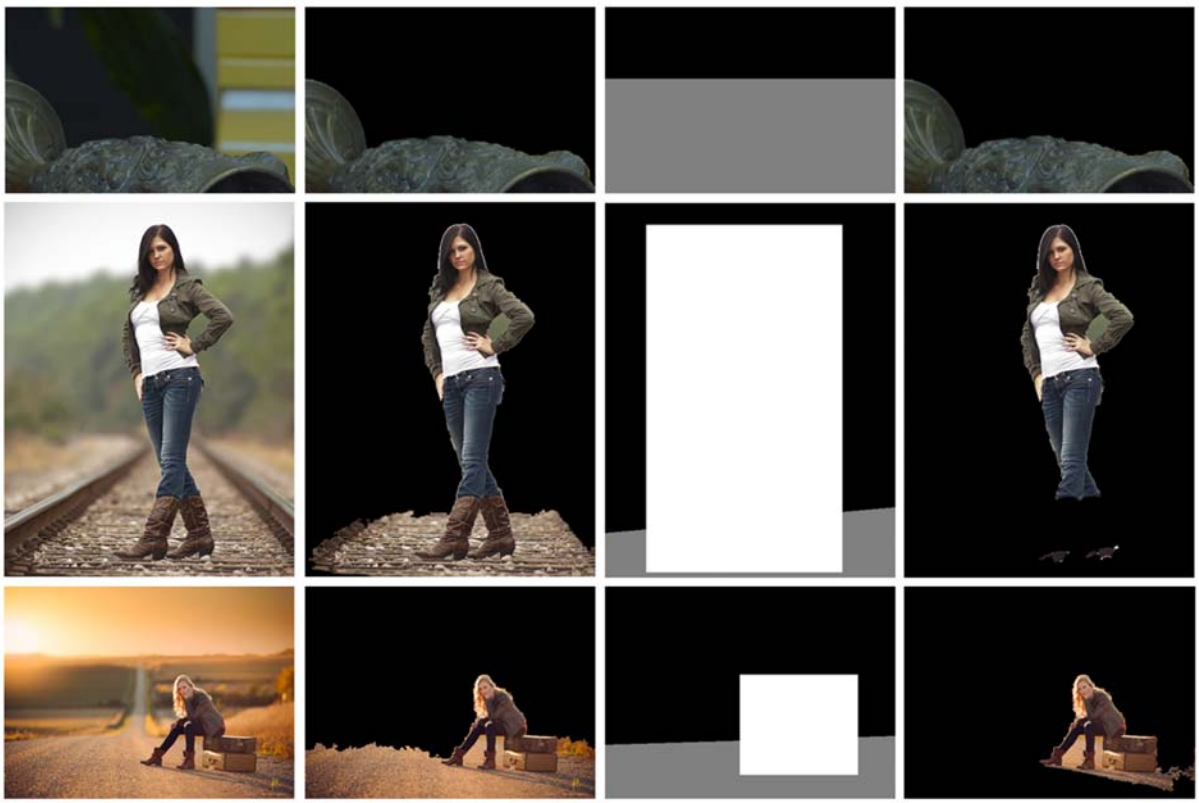


Fig. 37 LDOF image classification and OOI matting results: *Goblet*, *Girl-Tracks* and *Girl-Hitchhiker* (from top to bottom). (a) Original defocused image; (b) ROI; (c) Classification; (d) Matted OOI.

Another limitation of the proposed method is the high number of heuristic strategies employed. Further research is necessary to possibly amalgamate and refine this approach. The segmentation may also be affected when considering scenarios where the ground region is multi-layered or multi-coloured. Further research is warranted in this regard.

E. Conclusion

A novel unsupervised method is proposed for the accurate delineation or extraction of the OOI from a single LDOF image when the OOI is located within a larger ROI. Firstly, a baseline reference of the OOI is determined by using the dominant gradients in the image. Secondly, an approximation of the OOI is produced by expanding these baseline artefacts within the confines

of the ROI. Thirdly, the image is classified into one of six types of LDOF image scenarios. Fourthly, segmentation is performed using *k*-means clustering and, under the guidance of the OOI approximation, the non-OOI region are initially identified and subsequently excluded from the ROI thereby isolating only the closed boundary region of the OOI. This final stage is dependent on the type of LDOF image.

In an unsupervised OOI matting model no a priori reference for the OOI exists and as such the problem becomes ill-posed. However, the results show that the interrogation of the dominant gradients in a LDOF image is an extremely effective means of identifying the OOI and also providing a rudimentary or baseline reference for the approximation of the OOI. Matting is only applicable to two of the six proposed types of LDOF images. For the other four types the initial extraction of the ROI provides a sufficient representation of the OOI. In this research the initial extraction of the ROI is achieved using the method proposed in the previous section on *p*. 18 entitled *Unsupervised Region-of-Interest Extraction based on the Correlation of Gradient and Higher-Order Statistics Saliencies*.

In some scenarios the OOI may represent the entire ROI and in others the ROI may be comprised of two regions. These include the OOI and the non-OOI region. The non-OOI region are usually referred to as the ground-plane region; this is denoted by GOI. By this understanding the matting of the OOI essentially involves the removal of GOI from within the ROI. The identification and subsequent removal of GOI is performed in three stages. Firstly, segmentation of the regions within the LDOF image is performed. The results show that *k*-means clustering is adequately suited for LDOF images. Secondly, an approximation of the OOI is produced within the confines of the ROI by correlating the edges and cluster segments using the OOI baseline reference as a guide. Thirdly, an approximation of GOI is produced within the confines of the ROI by invalidating the cluster segments against the approximation of the OOI. The results show that this feedback approach is the most effective means of isolating the closed boundaries of the OOI from the rest of the ROI.

Image classifications plays a vital role in several arenas of image processing. In this research a novel method is proposed for the classification of LDOF images. The classification of a LDOF image is actually a classification of the ROI within the LDOF image and may be considered as being independent of the matting of the OOI. Six types of ROI scenarios are described in this research. The results show that the dominant gradients within the ROI provide an adequate reference for the OOI and the proposed expansion and refinement of this region within the confines of the ROI is provides an extremely robust means of determining the OOI-ROI permutation and subsequently the classification of the LDOF image into one of the six proposed types.

Future work may include the matting of the OOI as well as the classification of the LDOF image independent of the extraction of the ROI.

IV. AUTONOMOUS DEPTH MAP GENERATION OF SINGLE 2D LOW DEPTH-OF-FIELD IMAGES USING DEFOCUS, LINEAR PERSPECTIVE AND GESTALT PRINCIPLES

A. Introduction

A depth map may be described as a 2D representation of 3D space whereby every pixel in an image is given an associated relative depth value. This is usually in the form of a grey-level intensity array, as illustrated in Fig. 38. Depth maps play a vital role in arenas such as 3D computer graphics, machine vision and 2D-to-3D conversion.

Fig. 38 An input image and its estimated depth map [40]. In this case the higher the intensity of the pixel (whiter) implies a closer relative depth to the camera.

There is no optimal solution for the unsupervised generation of a depth map from a single 2D image. Even if the regions in a scene are accurately delineated there is still no precise means of assigning depth to these regions without a priori information such as the parameters of the capture device. The fundamental problem is that given only limited information in two dimensions, a third dimension, depth, has to be extrapolated.

There are two broad associated complexities with respect to depth map generation. The first involves the segmentation of the objects and regions within the image and the second entails the actual assignment of depth values to these objects and regions.

For the assignment of depth to be effective it is necessary to account for the multiple objects and sub-regions present within a scene; all of which have to be meticulously delineated and categorised. In addition to accurately delineating these objects and regions, objects effectively occluding (in front of and partially blocking) other objects need to be identified. This is essential, since every pixel has to be grouped and assigned an appropriate and relative depth value.

This images considered in this research are single 2D low depth-of-field images. In photography and cinematography, the term depth-of-field (DOF) refers to amount of focus (sharpness) or defocus (blurriness) present in an image. Image or video capture is broadly categorised into two DOF techniques. These include deep or high DOF and shallow or low DOF (LDOF). Illustrations of both these DOF methods are provided in Fig. 9.

Fig. 39 Depth-of-field [66]. (a) High DOF. (b) Low DOF.

The categorisation of segments of regions in an image is commonly referred to as labelling [43]. Some of the most common labels include the object of interest (OOI), ground region, background and skyline. Without any a priori baseline reference for the regions in an image the labelling of these regions may have to rely on Gestalt principles [24].

Emphasis in a LDOF image is usually placed on the in focus OOI. The precise delineation of this region, which is referred to as image matting [114], is crucial to the accuracy of the depth map, since all if not a significant majority of the pixels constituting the OOI, especially those at the boundaries, need to be accounted for prior to the assignment of depth.

This research proposes a model for the unsupervised estimation and allocation of relative depths across an entire single 2D LDOF image. This is achieved through the incorporation of LDOF image classification, OOI matting, Gestalt-based region classification, gradient-plane estimation [23, 56] using vanishing point detection [23, 56] and Gaussian blur analysis [42, 51, 53, 60, 63, 69, 82-84]. The final product is presented in the form of a grey-level intensity map.

This paper is organised as follows: Section B provides a brief discussion of some of the related work. Section C provides a description of the proposed model as well as the necessary steps required to autonomously generate a depth map from a single 2D LDOF image. Section D reports the experimental results. Section E provides a brief summary of the research and closes the paper.

B. Related Work

Depth from defocus (DfD) deals with how objects are perceived depending on their distance from the camera's focal plane based on the relationship between the focused and defocused regions in an image. [87]. In LDOF images objects will exhibit a certain amount of blur (out of focus) relative to their position very near to, or very far from, the object in focus. Fig. 40 shows three different images of the same scene shot with three different LDOF focal settings [115].

Fig. 40 Depth from defocus [115]. (a) Apple in focus; (b) Grapefruit in focus; (c) Background in focus.

In Fig. 40(a) the apple is in focus while the grapefruit and the BG appear out of focus. The BG appears more out of focus (defocused) than the grapefruit because it is further away from the apple. In Fig. 40(b) the grapefruit is brought into focus, with the apple and BG defocused. Although the BG is still defocussed it nevertheless exhibits much less blur than in Fig. 40(a), agreeing with the concept of distance from the focal plane. In Fig. 40(c) the BG is brought into focus and both the grapefruit and apple are defocused. Again it is observed that owing to its distance from the background the apple is more defocused than the grapefruit. These attributes may provide a means extrapolating the relative depths across an entire image.

Fig. 41 Defocus blur [116].

By comparing the two images shown in Fig. 41 the relationship between defocus and the optical setting of the camera may be further emphasised. In both these images LDOF settings are used, resulting in the OOI appearing in focus and the background appearing to be out of

focus. Apart from the minimal angular difference, the OOI appears to be sharp and approximately the same size in both images. At first glance the background in A.1 appears to be more blurred than the background in B.1. However, this is only true in reference to *absolute blur* and not *relative blur*. In terms of the latter there is a direct association with *magnification* [60]. When taking this into account, the depth of distant background objects always holds true in relation to the relative blur. Fig. 42 illustrates this concept, whereby both background objects are magnified to the same size and appear to have the same diameters of blur kernels with approximately equal Gaussian spreads.

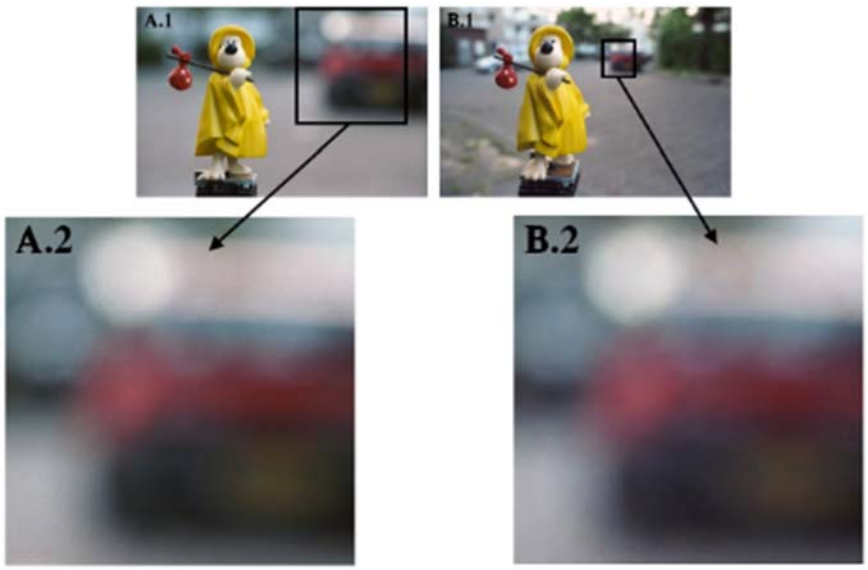


Fig. 42 Relative blur [116]

Most of the early work into DfD was done by Pentland [87], who introduced the concept of the blur circle (also known as circle of confusion) into the DfD framework. This research was closely followed by Grossman [117] and Subbarao [118, 119].

For a camera with lens of focal length[4] f_{L}, the relationship between the position of a point close to the optical axis in a scene and the position of its focused image is given by the familiar thin lens formula [120]:

$$\frac{1}{f_{\mathrm{L}}} = \frac{1}{u} + \frac{1}{v}, \tag{9}$$

[4] Depends on the shape and material of the lens. For the thin lens model $\frac{1}{f_{\mathrm{L}}} \approx (\frac{1}{R_1} - \frac{1}{R_2})(n - 1)$, where n is the refractive index of the lens and R_1 and R_2 are the radii of the two surfaces of the lens.

where u represents the distance between the in focus object (focal plane) and the lens and v, which is the *focus setting*, represents the distance between the lens and the image plane. This is illustrated in Fig. 43.

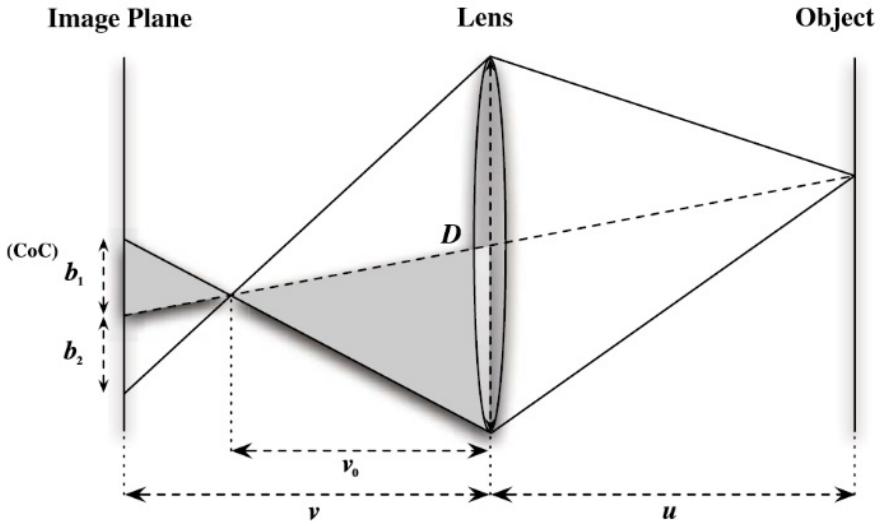


Fig. 43 Geometry of (real) finite-aperture imaging system [121]

Two scenarios exist when considering Eq. (9) and the associated variables given in Fig. 43; these include $v = v_0$ and $v \neq v_0$. For the former a point on an object is projected on the image plane as a point. Subsequently all points located at this focus distance (on the focal plane) will appear as points on the image plane. For the latter the thin lens law is violated. In this case a point from an object is projected on the image plane not as a point but rather as a disk (assuming a circular lens). This disk is referred to as the blur circle or circle of confusion (CoC). Geometrically, the equation for the radius of this disk, denoted by c, is given by

$$c = \frac{D}{2}\left|1 - \frac{v}{v_0}\right| \tag{10}$$

where D is the aperture of the lens.[5] It may also be shown that $b_1 = b_2 = c$. Moreover, by analysing the geometry of Fig. 43 the following equation may be obtained for the estimation the depth:

[5] D is not necessarily equal to the diameter of the lens since some cameras come equipped with a diaphragm, a mechanism that generates a variable-diameter opening in front of the lens. For standard cameras $D = \frac{f_L}{N}$, where N is referred to as either the f-number, f-stop or relative aperture. Some references use f_0 instead of N.

$$u = \frac{f_L v}{v - f_L - 2cN} \quad (\text{for } u > v)\,. \tag{11}$$

Pentland showed that this CoC is a nonlinear monotonic function relative to u and may be modelled as the convolution of a focused image with a *Gaussian* point spread function (PSF). If P is the real-world defocused point, then p at coordinates (x, y) is the projected image point having a blur radius of $\sigma(x, y)$. By letting $\sigma(x, y) = \sigma$, where σ is a constant for a given window, the PSF may then be expressed as

$$h(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right). \tag{12}$$

At this point an important conclusion is drawn: the width σ of the PSF $h(x, y)$, which is assumed to be of Gaussian shape, is proportional to the CoC or c, such that

$$\sigma = \gamma c, \tag{13}$$

where γ becomes a camera-specific proportionality constant. From this it is inferred that the depth may be determined by the PSF difference between a focused and a defocused image, hence the terminology “depth from defocus”.

Based on these observations it would seem that the direct interrogation of the Gaussian PSF would be the most appropriate method for defocus analysis. Although this is true it has proven to be a complicated research area, especially when considering blind or unsupervised approaches where no a priori knowledge is available.

The regions of defocus in an image may be visualised as being in focus areas having subsequently been convolved with a circular Gaussian PSF. This is described by the following linear model equations [94]:

$$i_b(x, y) = i_0(x, y) \otimes h(x, y) \qquad \text{(spatial domain)} \tag{14}$$

$$I_b(\mu, v) = I_0(\mu, v)\ \mathscr{H}(\mu, v) \qquad \text{(frequency domain)} \tag{15}$$

where i_b represents the defocused (blurred) image, i_0 represents the *true* image, h represents the impulse response or the PSF responsible for the blurring of i_0 (refer to Eq. (12)), and I_b, I_0 and $\mathscr{H}$ refer to the equivalent in the frequency domain, respectively; $\otimes$ denotes 2D convolution.

The reverse of the convolution process is termed *deconvolution*, the aim of which is to recover i_0 given i_b and h. Even with these variables known a priori, the solution to the

deconvolution problem, through inverse filtering, may nevertheless be counterproductive owing to the effects of noise amplification.

Further complications exist with deconvolution when both i_0 and i_b are not known. This situation is termed *blind* deconvolution. Solutions to this ill-posed problem have been proposed [122]. These usually begin with an initial guess of i_0, which is often taken to be the blurred image itself; subsequently, attempts are made to obtain h through a least squares solution; and finally, the iteration is reversed and attempts are made to estimate i_0 based on the estimated h. The disadvantage of these approaches is that they may prove to be computationally expensive, owing to the size of h usually being much smaller than i_0.

In order to improve efficiency, techniques such as *iterative* blind deconvolution (IBD) [123] have been proposed. In the case of IBD, the 2D FFT of i_0 and h are alternatively updated until $I_b(\mu, \nu) \cong I_0(\mu, \nu)\, \mathscr{H}(\mu, \nu)$ is almost satisfied; in this scenario the convergence progression, as well as robustness towards noise, is shown to be very much improved [122-124].

Deconvolution has applications in several research fields. However, for this technique to be viable in relative depth estimation, the PSF would have to be applied to each of the estimated defocused regions and not on the entire image as is currently proposed. Although attempts to improve the efficiency and reliability of image-based deconvolution have been put forward, such as the coded aperture method [125], these techniques nevertheless require prior physical modifications to the capture device, making them impractical for the purpose described in this research.

An alternative approach to the PSF technique is the reverse heat equation method. This technique, which was originally proposed by Gabor (1965) [126], may also be applied to the image de-blurring problem. Since the Gaussian PSF is a fundamental solution to the heat equation, the reverse of the PSF may effectively be formulated in terms of the reverse of the heat equation. Another study proposes that the reverse heat equation be iteratively applied to each of the edge pixels, with the point of convergence being determined at each step by the difference in the gradient of the edge pixel and the average gradient of the local neighbourhood [82]. This approximation of the convergence is referred to as the reverse diffusion time. This approach, which is proposed as an estimated measure of the relative depth, is termed the inverse diffusion method.

DfD, as described above, was originally designed to estimate depth by firstly, using two or more images and secondly, by having knowledge of the camera parameters. For single LDOF images two problems becomes apparent. Firstly, there is only a one image and secondly, the camera parameters are essentially unknown. This may seemingly make the DfD method impractical for unsupervised depth recovery. However, several DfD methods have been proposed for the extraction of relative depths from single LDOF images by analysing the blurred regions through manipulation of the gradient information [41, 42, 51, 61-63, 65, 67, 70, 82, 83, 127].

These blur analysis techniques, which are often referred to as Gaussian blurring or re-blurring methods, are based on the direct interrogation of the complete blur across the LDOF image.

As discussed above, the defocused regions of an image are defined as having characteristics that are Gaussian in nature and may be described as in focus areas of an image having been convolved with a circular Gaussian PSF. By analysing the gradient differences between a reference image and images of the same scene with different Gaussian blur attributes it may be possible to infer some degree of relative depth. The rationale is that across these images the objects in the focussed regions will contain *relatively* sharper gradients (fast high amplitude variations) than objects in the defocussed regions (small amplitude variations). The problem with single defocussed images is that these secondary images are non-existent.

A technique to resolve this discrepancy has been proposed whereby a second image may be created by applying a Gaussian filter to the original defocused source (reference) image. This *second* image, in relation to the reference image, will have the appearance of an increased blur (smoothness) across the entire image. The approximate relative depths may then be estimated by calculating the differences between the reference image and the new *re-blurred* image. LDOF images by definition contain a certain amount of blur, therefore, the term re-blurred as it is used here actually implies the additional blurring of an image containing some initial amount of blur.

Several methods exploiting this technique have been proposed for the delineation of the ROI in LDOF images [42, 51]. In these cases a type of focused saliency map (FSM) is employed. The FSM, which effectively emphasises the focused regions in the image, is produced by performing a direct subtraction of the aforementioned re-blurred image from the source image.

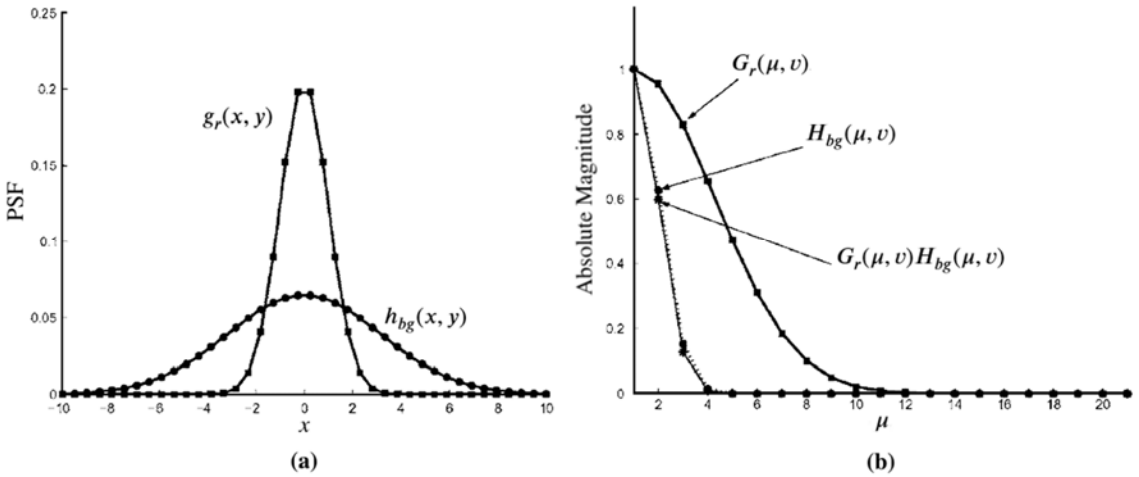


Fig. 44 Gaussian Blur [42]. (a) Two PSF with different blur parameters ($\sigma_{hbg} = 3.16$ and $\sigma_{gr} = 1$). (b) Corresponding Fourier transforms.

A defocused image $i_{df}(x, y)$ may be described as comprising two components. These include the FG, $i_{fg}(x, y)$, and the BG, $i_{bg}(x, y)$. The latter is described as a region containing a non-zero

PSF, compared to the approximate zero PSF of the focused FG region; this original non-zero PSF is denoted by $h_{bg}(x, y)$. To generate the afore discussed second image, $i_r(x, y)$, a PSF is applied to $i_{df}(x, y)$. This is described by the following equation:

$$i_r(x, y) = i_{df}(x, y) \otimes g_r(x, y) , \tag{16}$$

where $g_r(x, y)$ denotes the re-blurring PSF and $\otimes$ denotes convolution.

If G_r and H_{bg} refer to the Fourier transform of g_r and h_{bg}, respectively, then the product of these two PSFs may be expressed as

$$G_r(\mu, \upsilon)H_{bg}(\mu, \upsilon) = \exp(-2\pi^2\sigma_{hbg}^2 \left(1 + \left(\frac{\sigma_{gr}}{\sigma_{hbg}}\right)^2\right)(\mu^2 + \upsilon^2) . \tag{17}$$

From the Eq. (17), if $\sigma_{gr} < \sigma_{hbg}$, then the value of $1 + \left(\frac{\sigma_{gr}}{\sigma_{hbg}}\right)^2$ will tend towards 1, resulting in $G_r(\mu, \upsilon)H_{bg}(\mu, \upsilon) \cong H_{bg}(\mu, \upsilon)$. For LDOF images it is assumed that most of the blurred regions will have parameters $\sigma_{hbg} > 1$. By setting σ_{hbg}= 3.16 (larger) and $\sigma_{gr} = 1$ (smaller), the curve of $G_r(\mu, \upsilon)H_{bg}(\mu, \upsilon)$ is almost identical to $H_{bg}(\mu, \upsilon)$, as illustrated in Fig. 44.

Based on the outcome displayed in Fig. 44(b) it may be concluded that $I_{df}(\mu, \upsilon) - I_r(\mu, \upsilon)$ will result in most of the energy corresponding to the focused area in the image; the FSM of a defocused image is then defined as

$$\text{FSM}(x, y) = \left|i_{df}(x, y) - i_r(x, y)\right| . \tag{18}$$

Owing to noise and soft shadows, some inconsistencies may arise in the FSM. To attenuate some of these isolated points a smoothing filter is usually applied to the FSM. The extraction of the ROI is performed using a three-step process. Firstly, a trimap is constructed by applying morphological filtering and thresholding to the FSM [41]. This trimap comprises of three regions viz. definite FG regions, definite BG regions and unknown border regions. Secondly, an initial estimate of the ROI is made by excluding the definite BG regions from the trimap. Thirdly, the ROI is refining by applying adaptive error control matting [84] to the unknown border regions. Although the FSM method described here may be used to address the ROI matting problem it is nevertheless insufficient for extrapolating image-wide relative depth information.

Other methods, based on the analysis of the Gaussian blur, have been proposed for estimating the relative depths of entire single defocused images. Instead of subtracting the second re-blurred image from the original source image [42], the defocus blur amount is estimated by obtaining the gradient ratio between the original source image and the second re-blurred image [63]. The defocus blur is initially estimated at edge locations and then propagated across the entire image,

producing a relative full depth map. This method is closely related to the earlier discussed inverse diffusion method [82].

By assuming the edges in an image are *ideal* step edges, the following function may be considered:

$$i_e(x) = Au(x) + B, \tag{19}$$

where $u(x)$ is the step function, A is the amplitude, and B is the offset of the edge located at $x = 0$. However, the edges in a defocused image are not truly ideal and therefore contain some degree of blurring; this is described as

$$i_{e0}(x) = i_e(x) \otimes g_0\left(x, \sigma_{g0}\right), \tag{20}$$

where $i_{e0}(x)$ represents the edges of the defocused image and $g_0\left(x, \sigma_{g0}\right)$ represents the Gaussian PSF where the standard deviation, σ_{g0}, is specific to each of the edge pixels. The ratio between the gradient of the original step edges and the defocused (blurred) edges may then be calculated at these edge locations.

For convenience the 1D case is first described and then extended to 2D. Based on Eq. (19) and Eq. (20) it is shown that

$$\begin{aligned}
\nabla_{i_{er}}(x) &= \nabla\left(i_{e0}(x) \otimes g_r\left(x, \sigma_{gr}\right)\right) \\
&= \nabla\left((Au(x) + B) \otimes g_0\left(x, \sigma_{g0}\right) \otimes g_r\left(x, \sigma_{gr}\right)\right) \\
&= \frac{A}{\sqrt{2\pi\left(\sigma_{g0}^2 + \sigma_{gr}^2\right)}} \exp\left(-\frac{x^2}{2\left(\sigma_{g0}^2 + \sigma_{gr}^2\right)}\right),
\end{aligned} \tag{21}$$

where σ_{gr} is the standard deviation of the re-blurring Gaussian kernel, $g_r\left(x, \sigma_{gr}\right)$. The magnitude of the gradient ratio between the original defocused source image and the re-blurred second image is then described as

$$\frac{\left|\nabla_{i_{e0}}(x)\right|}{\left|\nabla_{i_{er}}(x)\right|} = \sqrt{\frac{\sigma_{g0}^2 + \sigma_{gr}^2}{\sigma_{g0}^2}} \exp\left(\frac{x^2}{2\sigma_{g0}^2} - \frac{x^2}{2\left(\sigma_{g0}^2 + \sigma_{gr}^2\right)}\right). \tag{22}$$

The gradient ratio is actually maximum at the edge locations, which are located at $x = 0$. This then gives the maximum ratio as

$$R = \frac{\left|\nabla_{i_{e0}}(x)\right|}{\left|\nabla_{i_{er}}(x)\right|} = \sqrt{\frac{\sigma_{g0}^2 + \sigma_{gr}^2}{\sigma_{g0}^2}}. \tag{23}$$

From above, the maximum ratio is dependent on only σ_{g0} and σ_{gr}. Rearranging Eq. (23), the defocus blur may then be defined as

$$\sigma_{g0} = \frac{1}{\sqrt{R^2 - 1}}\ \sigma_{gr}. \tag{24}$$

The proposed blur estimation method is illustrated in Fig. 45.

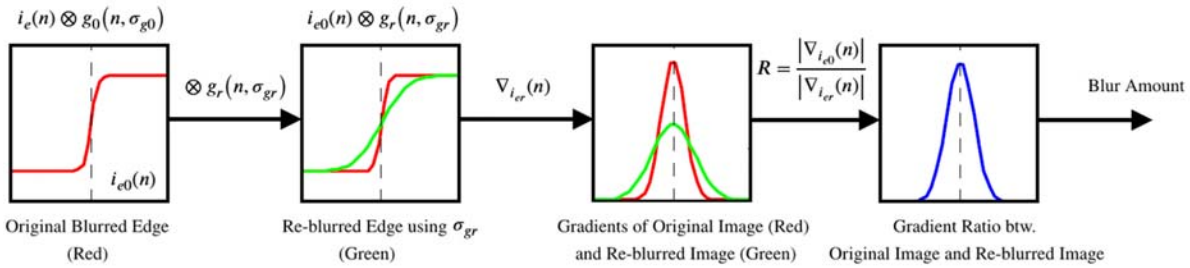


Fig. 45 Blur Estimation (Image adapted from [63]).

The 1D model, which is described above, is extended to 2D by using

$$\|\nabla_i(x, y)\| = \sqrt{\nabla_{i_x}^2 + \nabla_{i_y}^2}\ , \tag{25}$$

where ∇_{i_x} and ∇_{i_y} are the gradients in the x and y directions, respectively. By setting $\sigma_{gr} = 1$ and determining the edge locations by the Canny method [95], the sparse defocus (depth) map of a single defocused image is then defined as

$$\widehat{dm}(x, y) = \frac{1}{\sqrt{\left(\frac{\left|\nabla_{i_{e0}}(x,y)\right|}{\left|\nabla_{i_{er}}(x,y)\right|}\right)^2 - 1}}. \tag{26}$$

To minimise some of the inconsistencies, a bilateral filter (BF) [128] is applied to the sparse saliency map, $\widehat{dm}(x, y)$. A final defocus map $dm(x, y)$ is produced by interpolating the sparse saliency map values at the edge locations across the entire image.

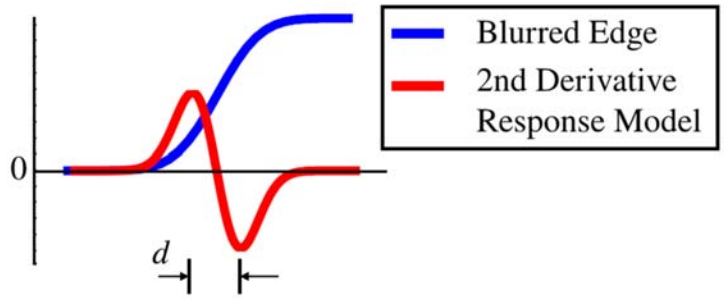


Fig. 46 Model for distance between second-derivative extrema [60]

Another proposed approach to the Gaussian re-blurring technique uses the second derivative Gaussian filter response as a means to determine the defocus blur of an image [60, 127]; this is illustrated in Fig. 46. Any given edge in an image is typically a non-ideal step edge i.e. it is described as either an error function or as an ideal step edge that is blurred by a Gaussian, as described by Eq. (20) on *p*. 86. Furthermore, the derivative of Eq. (20) will actually yield a Gaussian, and the second derivative of Eq. (20), which equates to the first derivative of a Gaussian, will produce a zero-crossing response as illustrated in Fig. 46.

By adapting Eq. (20) on *p*. 86, the second derivative filter is modelled by

$$\frac{\partial}{\partial x^2} i_{e0}(x,y) = \frac{-Ax}{\sqrt{2\pi\left(\sigma_{g0}^2 + \sigma_{gr}^2\right)^3}} \exp\left(-\frac{x^2}{2\left(\sigma_{g0}^2 + \sigma_{gr}^2\right)}\right), \tag{27}$$

$$= \frac{-Ax}{\sqrt{2\pi}\left(\frac{d}{2}\right)^3} \exp\left(-\frac{x^2}{2\left(\frac{d}{2}\right)^2}\right), \tag{28}$$

where

$$\left(\frac{d}{2}\right)^2 = \sigma_{g0}^2 + \sigma_{gr}^2 \,. \tag{29}$$

Rearranging Eq. (29) then gives

$$\sigma_{g0} = \sqrt{\left(\frac{d}{2}\right)^2 - \sigma_{gr}^2} \,, \tag{30}$$

where d is the measure of the distance between the extrema, as illustrated in Fig. 46, σ_{gr} is the pre-defined re-blur kernel and σ_{g0} represents the computed estimate of the blur kernels at edge locations of the original defocused source image [60].

Owing to factors such as noise, soft shadows, glossy highlights and intensity consistent regions, some irregular estimates of the defocus blur may be present in the sparse defocus map. As with the previous approach, these inconsistencies or outliers may be somewhat attenuated through the application of some type of edge-preserving smoothing filter, such as a bilateral filter [128, 129]. A full defocus map is subsequently produced by interpolating the edge values of the filtered sparse defocus map across the entire image using, for example, a local neighbourhood pixel intensity [130], matting Laplacian [45] or morphological filtering and thresholding [42]. Although the defocus blur maps described here are not equivalent to relative depth maps, they may nevertheless be considered as rudimentary estimates of the relative depths.

C. Proposed Approach

The proposed model involves five stages. The first stage considers the classification of the LDOF image. The second stage involves the delineation or matting of the OOI from the rest of the image. Both these stages are described in detail in the previous section *Unsupervised Matting of the Object-of-Interest in Low Depth-of-Field Images* on *p*. 54. The third stage determines the method employed for the assignment of relative depths across the image. This is dependent on the classification of the LDOF image as well as the description of the non-OOI region. In the proposed model two descriptors are considered; these include equidistant- and gradient-plane-based. The fourth stage is dependent on the outcome of the third stage. For the equidistant-based approach the defocussed regions across the entire image are interrogated using Gaussian blur analysis. For the gradient-plane-based approach the non-OOI region is initially segmented and subsequently labelled according to Gestalt principles. The fifth and final stage, which is also dependent on the outcome of the third stage, involves the generation of the depth map by assigning relative depth values across the entire image.

For the equidistant-based approach the OOI is initially considered as the region of shallowest depth. Subsequently, the rest of the image is assigned relative depths based on the Gaussian blur analysis. For the gradient-plane-based approach a gradient-plane is initially produced by considering the vanishing point in the image. Subsequently, the relative depths are assigned to the Gestalt regions depending on their label and location along the gradient-plane.

1) Relative Depth Descriptor

The classification of a LDOF image is dependent on the depth-of-field settings of the capture device as well as the ROI and the position of the OOI. This is described and discussed in the previous section *Unsupervised Matting of the Object-of-Interest in Low Depth-of-Field Images*

on *p*. 54. In the proposed model six classification types are considered. These may be described using the illustrations provided in Fig. 47.

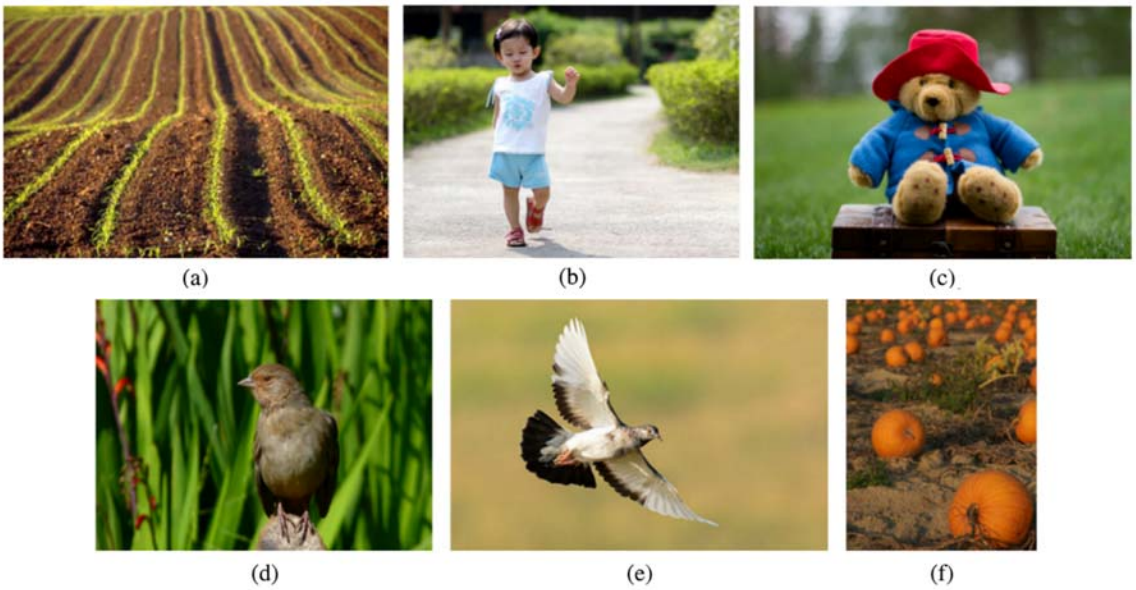


Fig. 47 Low depth-of-field images [60, 63]

Type 1 classification (refer to Fig. 47 (a)) contains no discernible OOI. Type 2, 3 and 6 classifications (refer to Fig. 47 (b), (c) and (f)) contain an OOI as well as an in focus ground region i.e. the OOI is a subcomponent of the ROI. For type 4 and 5 classifications (refer to Fig. 47 (d) and (e)) the OOI represents the entire ROI.

Depending on the classification of the image there may be two descriptions for the non-OOI region and consequently two approaches to the way the relative depth values may be assigned across the LDOF image.

The first is referred to as equidistant-based. This describes scenarios where all objects and sub-regions within the non-OOI region of the image are considered as being approximately equal distant behind the OOI. This descriptor is only applicable to type 4 and 5 classifications (refer to Fig. 47 (d) and (e)).

The second is referred to as gradient-plane-based. This describes scenarios where the objects and sub-regions in the non-OOI region of the image exhibit different depths relative to the OOI. This descriptor may apply to all classification types. Some of these regions may include the ground, which typically exhibits incremental relative depths, as well as objects (including the OOI) which lie at different locations on top of the ground (refer to Fig. 47 (a)–(c) and (f)). Another common region is the sky or some type of uniform area in the background (refer to Fig. 47 (b) and Fig. 47 (c)). For these scenarios the region in its entirety must be treated as being the furthest or deepest point in the image.

To determine which of these two scenarios describe the non-OOI region in a LDOF image, the proposed model considers firstly, segmenting the non-OOI region and secondly, grouping the sub-segments into ground, sky and unknown regions according to Gestalt principles and thirdly, interrogating the attributes of the ground and sky regions.

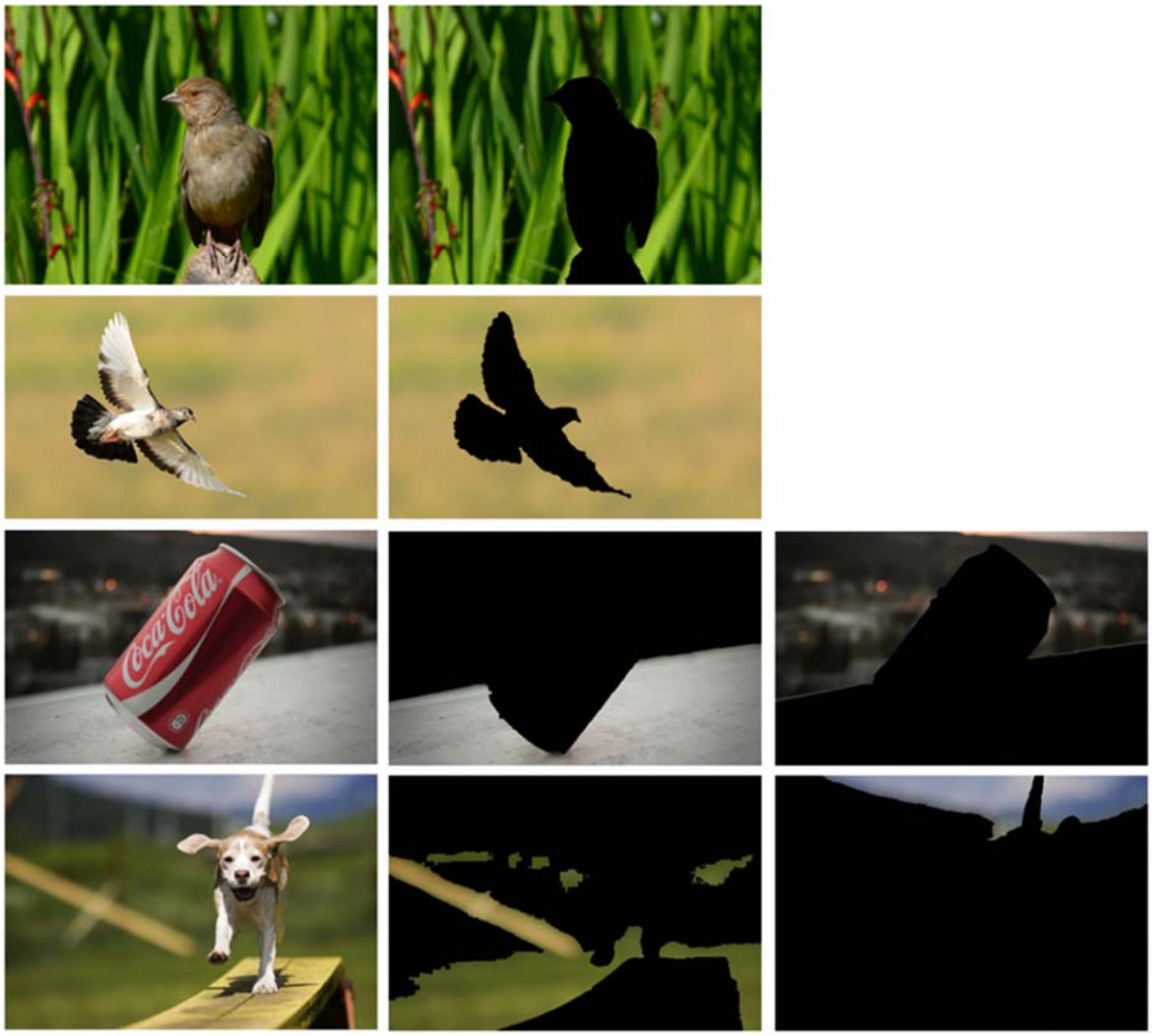


Fig. 48 Sub-classification of Type 4 and 5 classifications: *Bird-Reed*, *Bird-Flight*, *Coke* and *Dog* (from top to bottom) (a) Original defocused image; Extrapolated Gestalt-based (b) ground region and (c) sky region.

Two conditions have to be met for the image to be sub-classified as equidistant-based. Firstly, the LDOF image has to be classified as either type 4 or 5 and secondly, from experiment, either more than 30% of the perimeter of the ground region intersects with both the bottom row of the image as well as the non-OOI region or less than 30% of the perimeter of the sky region intersects with both the top row of the image as well as non-OOI region. If these conditions are not met, then by default the image is sub-classified as gradient-plane-based. Some type 4 and 5 classifications together with their Gestalt ground and sky regions are illustrated in Fig. 48. The *Bird-Reed* and

Bird-Flight images are classified as equidistant-based and the *Coke* and *Dog* images are classified as gradient-plane-based.

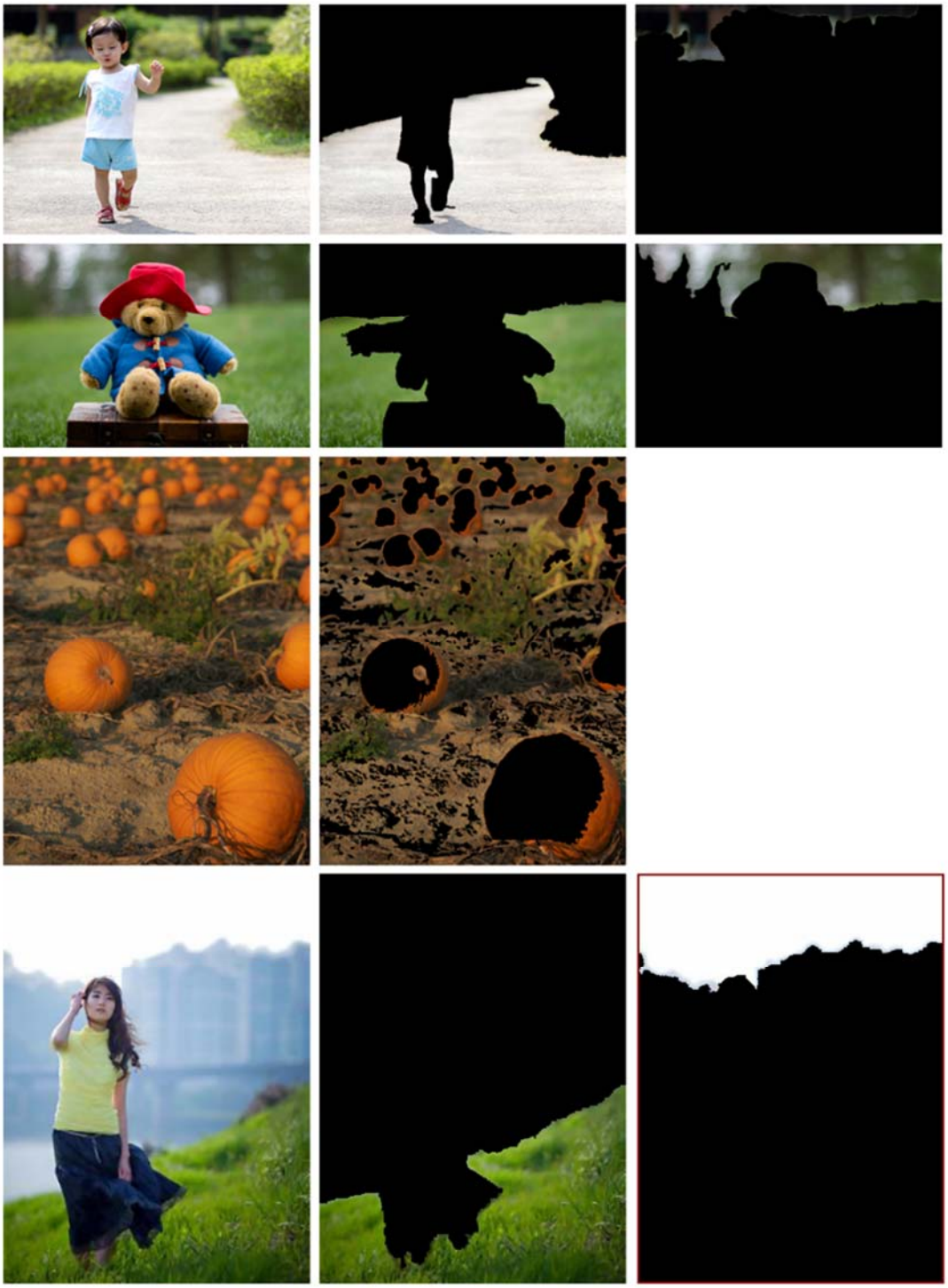


Fig. 49 Gradient-plane sub-classification. (a) Original defocused image; Extrapolated Gestalt-based (b) ground region and (c) sky region.

In the proposed model the non-OOI region is segmented using both *k*-means and binary quantisation clustering. The *k*-means segmentation technique is discussed on *p*. 35. Owing to the Gaussian spread nature of LDOF images certain significant boundary regions may become diluted. Although binary quantisation clustering is less effective than the *k*-means technique, when applied to the global segmentation of LDOF images, it is nevertheless observed from experiment to be more effective in dealing with diluted boundaries.

Binary quantisation is whereby a colour palette is optimally designed and each pixel is then assigned to a cluster based on the total squared error between the original pixels and the colours assigned from the palette [131]. The method chosen for this research is the one proposed by Orchard and Bouman [104]. In this case the colour palette is designed using a hierarchical binary tree structure where, based on the cluster covariance, eigenvalues, eigenvectors and mean distance, together with an erosion-based weighting, the pixels are separated according to their closeness to a plane perpendicular to the principal eigenvector and passing through the mean. Each original pixel is subsequently mapped to a cluster that is based on the closest distance to a colour in the palette using ordered dithering and error diffusion techniques.

From experiment, a cluster value of 3 is chosen for both the *k*-means and binary quantisation segmentation of the non-OOI region. Subsequently, ground and sky regions are produced separately per mode of segmentation. Using Gestalt principles, it is reasonable to assume that the ground region is located on the lower part of the image and the sky region is located on the upper part of the image. Based on this understanding and from experiment, the proposed model considers an array to be associated with the ground region if more than 30% of the perimeter of the entire cluster intersects with the bottom row of the image as well as non-OOI region and with the sky region if more than 30% of the perimeter of the entire cluster intersects with the top row of the image as well as non-OOI region.

Unsupervised systems are usually absent of any a priori information and consequently result in the problem becoming ill-posed. To account for these constraints, the proposed model correlates the *k*-means-based regions with the binary quantisation-based regions by intersecting the data.

For type 1 and 6 classifications the union, instead of the intersection, of the respective regions is considered. For the Gestalt ground region, the *k*-means-based ground region is intersected with binary quantisation-based ground region. Similarly, for the Gestalt sky region, the *k*-means-based sky region is intersected with binary quantisation-based sky region. Subsequently, any regions in the Gestalt sky region having an intersection overlap with the Gestalt ground region are considered to belong to the Gestalt ground region and excluded from the Gestalt sky region. Finally, any connected component not incident with top border or incident with the top and bottom borders are excluded from the Gestalt sky region. At this stage, for the purpose of classification, the non-OOI regions are broadly delineated into two Gestalt regions viz. ground and sky. However, five Gestalt sub-regions are actually considered when assigning relative depth values. This is discussed in the

following section on *p*. 95 and illustrated in Fig. 55 on *p*. 101. Examples of the Gestalt ground and sky regions are illustrated in Fig. 48 on *p*. 91 and Fig. 49 above.

2) Relative Depth Assignment

i. Equidistant-based

An image captured by a camera using a standard convex lens with a LDOF setting will contain objects at a particular distance from the camera that are in focus, as well as objects that are out of focus, to varying degrees, depending on their distance in front of or behind the focal plane. Based on this understanding it may therefore be possible to extrapolate a certain amount of information relating to the relative depths of the regions in a scene.

For the equidistant-based depth map the OOI is initially considered as the region of shallowest depth. Subsequently, the rest of the image is assigned relative depths based on the Gaussian blur analysis or more specifically the Gaussian re-blurred saliencies in the image. Some examples of these depth maps are illustrated in Fig. 50(b).

Fig. 50 Depth maps for type 4 and 5 classifications. (a) Original images; (b) Initial depth maps; (c) Finalised depth maps after application of bilateral filter.

For the Gaussian re-blurred saliency map the defocus blur amount is estimated through examination of the gradient ratios between the original image and a second blurred version of the original image; the so-called re-blurred image.

A sparse defocus map is produced by implementing the method proposed by Zhuo and Sim [63]. This is discussed on *p*. 85. To generate the re-blurred image a Gaussian low-pass filter is applied to a grayscale transformed version of the source image. From experiment a window size of 0.4% of the perimeter of the image with a sigma value of 1 is chosen. The gradients of both the

original image as well as the re-blurred image are determined by using a Sobel mask. In the proposed model any pixel not associated with the edge saliency map (defined on *p*. 26) is set to zero and to account for the full grayscale spectrum, the pixel intensities are extrapolated and distributed across the range [0, 255].

Some irregular estimates of the defocus blur may occur in the sparse defocus map. This is owing, in the main, to factors such as noise, soft shadows, glossy highlights and intensity consistent regions. Edge-preserving filtering is often used to attenuate some of these isolated points, spurious edges and outlier regions.

Tomasi and Manduchi (1998) [128] introduced the bilateral edge-preserving smoothing filter as an alternative to Perona and Malik's (1990) anisotropic diffusion method [132]. The former is a nonlinear filter functioning non-iteratively by weight averaging the colours of the neighbour pixels based on their distance in space and range. The latter is modelled on an iterative process employing partial differential equations [133].

The bilateral filtering of a pixel (x, y) is expressed as

$$\begin{aligned}\mathrm{BF}\big(\mathrm{FSM}(x,y)\big) = \\ \frac{1}{K}\sum_{(j,k)\in\eta(x,y)} \exp\left(-\frac{(j-x)^2+(k-y)^2}{2\sigma_s^2}\right) \cdot \\ \exp\left(-\frac{\big(\mathrm{FSM}(j,k)-\mathrm{FSM}(x,y)\big)^2}{2\sigma_r^2}\right)\mathrm{FSM}(j,k)\end{aligned} \tag{31}$$

where η(x,y) denotes a smoothing window centred at pixel (x, y), σ_s is the spatial Gaussian weighting that decreases the influence of pixels based on distance, σ_r is the range Gaussian weighting that decreases the influence of pixels based on intensity and K is the normalisation operator given by

$$K = \sum_{(j,k)\in\eta(x,y)} \exp\left(-\frac{(j-x)^2+(k-y)^2}{2\sigma_s^2}\right) \cdot \exp\left(-\frac{\big(\mathrm{FSM}(j,k)-\mathrm{FSM}(x,y)\big)^2}{2\sigma_r^2}\right). \tag{32}$$

Owing to the nonlinearity of the averaging process the bilateral filter (BF) becomes computationally intensive. As a consequence improvements to these type of filter methods have been proposed. In this research two of these modified filters are considered for the refinement of the Gaussian re-blurred saliency map. The first is a shiftable edge-preserving BF proposed by Chaudhury et. al [129, 134, 135]. In this case the exploitation of the *shiftability* associated with use of trigonometric range kernels allows for the realisation of the BF in constant time. The second is a domain transformed edge-preserving BF proposed by Gastal and Oliveira [133]. In this case

the original RGB image, which is represented by five dimensions viz. x, y, r, g and b, is frequency domain transformed to a lower dimension prior to filtering, while preserving distances. This concept is illustrated in Fig. 51.

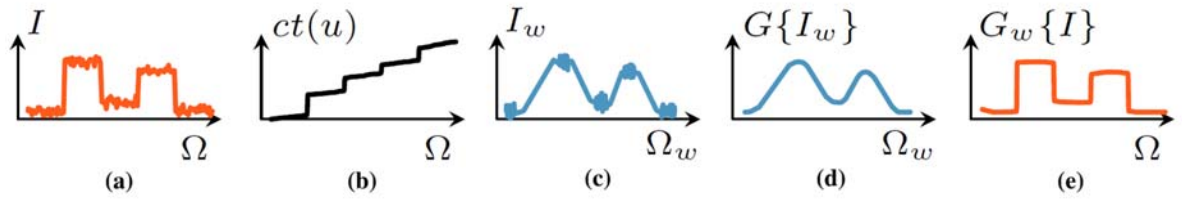


Fig. 51 Domain transform edge-aware filtering [133]. (a) Input signal I in original domain; (b) Associated domain transform; $ct(u)$; (c) I in the transformed domain; (d) Filtering using a Gaussian filter in the frequency domain; and (e) Filtered signal obtained through the reversal of $ct(u)$ for $G\{I_w\}$.

The following equation describes this filtering process:

$$J(p) = \int_{\Omega} I(p)F(\hat{p},\hat{q})\ dq = \int_{\Omega} I(p)H\left(t(\hat{p}),t(\hat{q})\right)\ dq \tag{33}$$

where I denotes the input image, F represents the BF in the spatial domain and H represents the filter kernel in the frequency domain. The resultant J denotes the output image derived through the filtering of I with either F or H, with

$$H\left(t(\hat{p}),t(\hat{q})\right) = \delta\{|t(\hat{p}) - t(\hat{q})| \leq r\}, \tag{34}$$

$$t(\hat{u}) = t\left(u, I(u)\right) = ct(u) = \int_0^u 1 + \frac{\sigma_s}{\sigma_r}\sum_{k=1}^{c}\left|I_k'(x)\right|\ dx, \tag{35}$$

and

$$r = \sigma_{H_i} = \sigma_H\sqrt{3}\times\left(\frac{2^{N-i}}{\sqrt{4^N-1}}\right), \tag{36}$$

where σ_s and σ_r refers to the spatial and range standard deviation of the signal, respectively, I_k is the k^{th} channel of I (for an RGB image, $c = 3$), σ_{H_i} is the standard deviation for the kernel used in the i^{th} iteration, N is the total number of iterations and $\sigma_H = \sigma_s$ (i.e. since variances and not standard deviations are added for each iteration, σ_{H_i} must halve at each iteration and have a squared sum matching the original desired variance, σ_H^2). The distance between neighbouring pixels in the spatial domain is typically computed using the L_2 norm. However, for the frequency domain the L_1 norm is considered. An added benefit of the domain transform edge-preserving

filter method is that the domain transformation and reversal through $ct(u)$ allows for the preservation of the geodesic distance between points on the curves, which adaptively warps the input signal and thereby allows for linear time edge-preserving filtering.

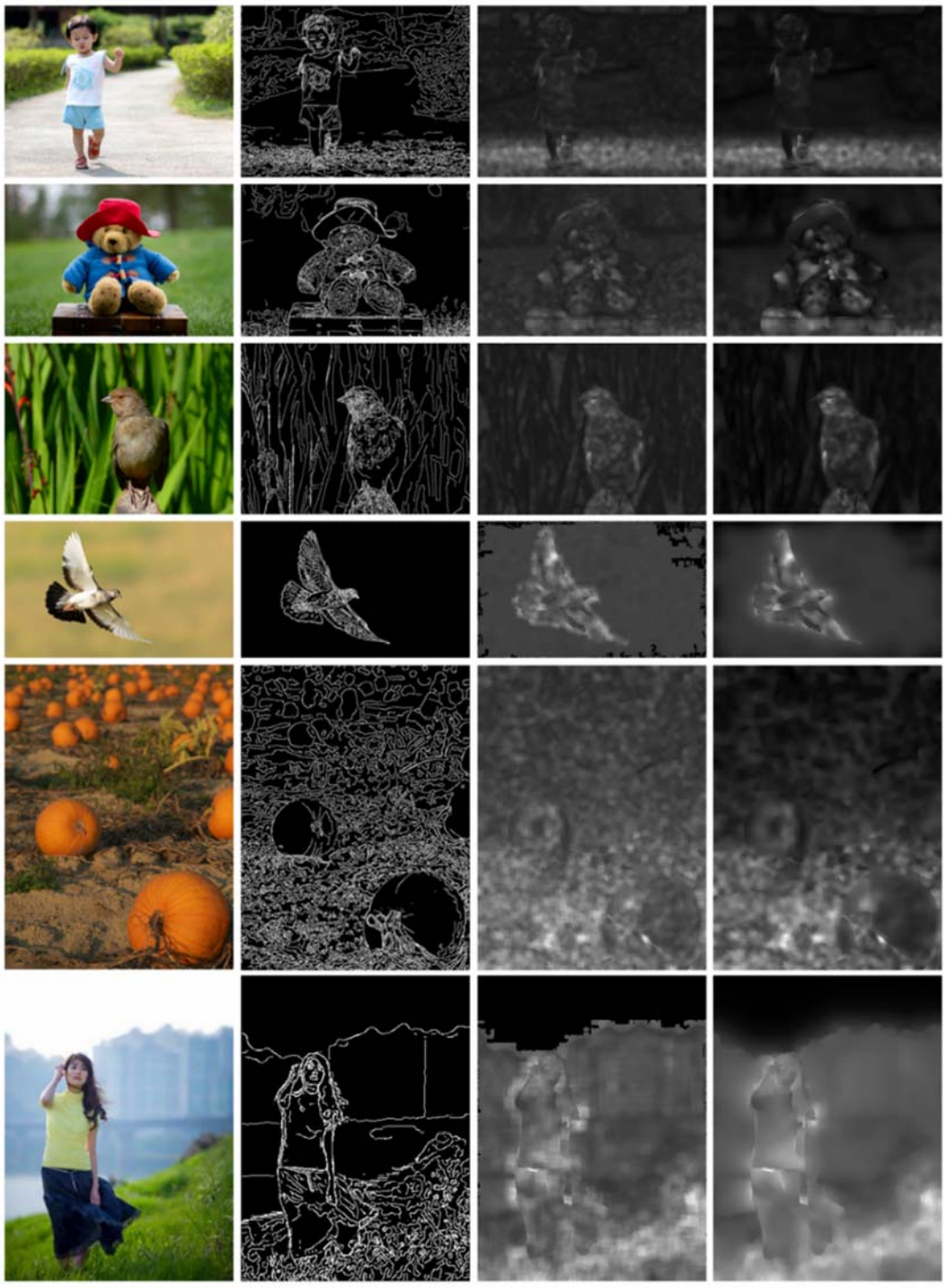


Fig. 52 Gaussian re-blurring saliencies. (a) Original images; (b) Edge saliency maps. (c) Defocus maps after application of first bilateral filter; (d) Defocus maps after application of second bilateral filter.

From experiment, a spatial standard deviation $\sigma_s = 0.65\%$ the perimeter of the image, a range standard deviation $\sigma_r = 1.7\sigma_s$ and a $n \times n$ smoothing window with $n = \sigma_s$ is chosen for the BF. For the domain transform edge-preserving filter a spatial standard deviation $\sigma_s = 0.9\%$ the perimeter of the image, the range standard deviation is set to 0.1 and the number of iterations is set to 3. The results of the two-stage filtering are illustrated in Fig. 52(c) and (d).

The equidistant-based depth map is produced by initially using the full smoothed Gaussian re-blur map as the foundation. Subsequently, the pixel intensities are distributed across the range [0, 255] with 255 representing the furthest distance from the observer. Finally, the pixel intensities of the OOI region are assigned a value of 0. Examples of this depth map are illustrated in Fig. 50(b) on *p*. 94.

ii. Gradient-Plane-based

For the gradient-plane-based depth map the objects and regions vary in relative depth across the image. In the proposed model the autonomous assigning of these relative depths is achieved in a three-staged approach. Initially, the non-OOI region is segmented, grouped and labelled into five Gestalt-based regions. For type 1 and 6 classifications the entire image is considered as the non-OOI region. Subsequently, a gradient-plane is produced by considering the vanishing point in the image. Finally, the depth map is generated by assigning relative depth values to the Gestalt-based regions according to their associated label as well as location along the gradient-plane. Some examples of these depth maps are illustrated in Fig. 53(b) on *p.* 99.

In the proposed model the five Gestalt-regions considered are the ground, sky, objects on the periphery of the ground, objects on top of the ground region and objects within the sky region. The ground and sky regions are discussed earlier on *p*. 91. Objects on the periphery of the ground are considered to be any non-sky regions that are external to the ground regions.

To determine the objects on top of the ground, the ground region is segmented using *k*-means clustering. From experiment, a cluster value of 3 is chosen. Initially, a base ground region is chosen. This is assumed, with reasonably high probability, to be the cluster array containing the most non-zero elements. Subsequently, the connected components in the remaining two cluster arrays are analysed and if it is not incident on either the outer perimeter of the ground region or the bottom border of the image, then it is considered to be a valid internal ground region object.

The same concept and process used to determine the ground objects are applied to the objects within the sky region. The difference is that a connected component is considered to be a valid internal sky region object only if it is not incident on the outer perimeter of the sky region.

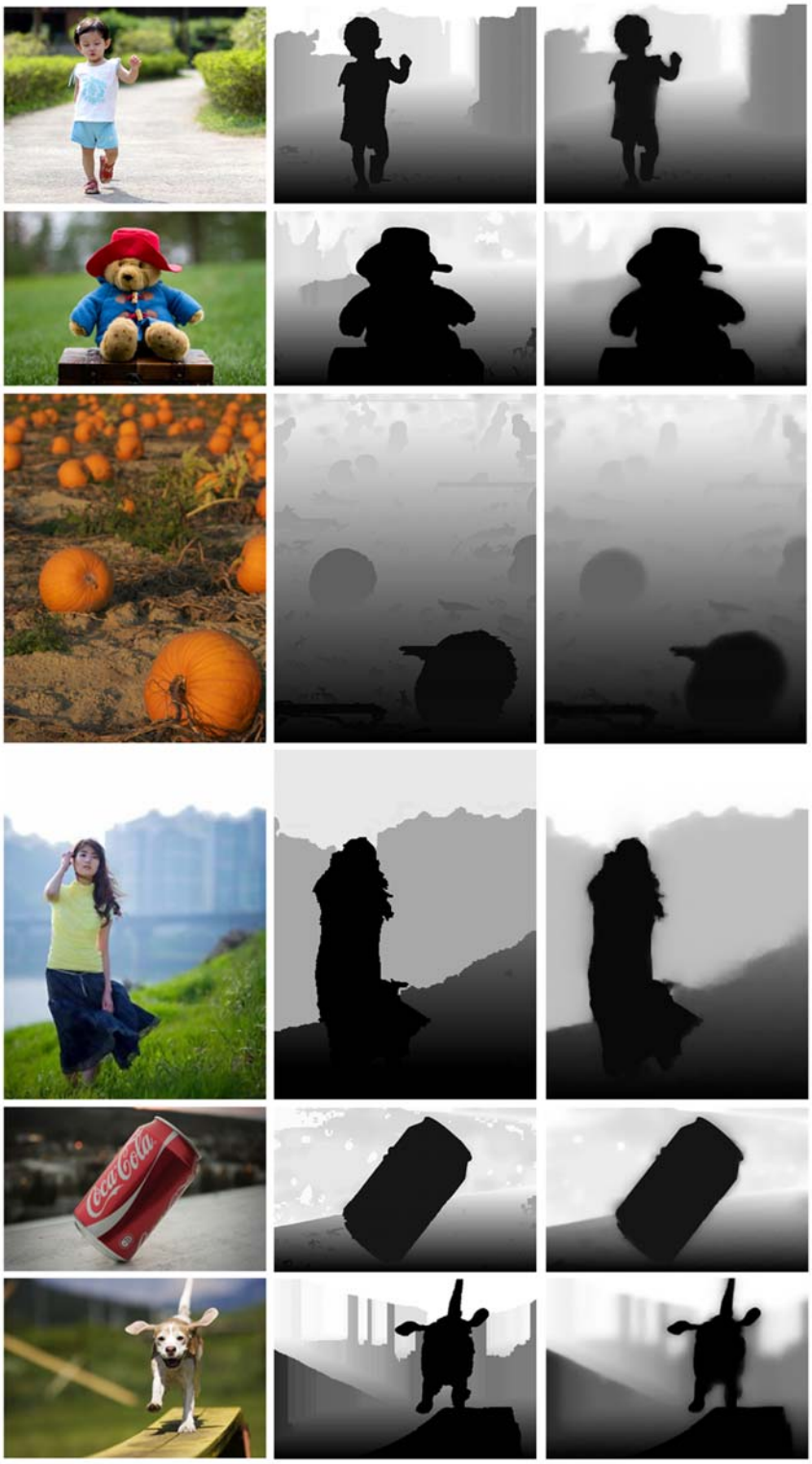


Fig. 53 Gradient-based depth maps: *Kid*, *Teddy*, *Pumpkin*, *Girl-City*, *Coke* and *Dog* (from top to bottom). (a) Original images; (b) Initial depth maps; (c) Finalised depth maps after application of bilateral filter.

A shortcoming with the proposed approach, in terms of defining the internal Gestalt regions, is that objects such as rocks overlapping with the bottom of the image and clouds overlapping with the top will not be considered as valid internal ground and sky objects, respectively. Although the assignment of relative depths to these objects may be incorrect, depending on their size, it may not necessarily have a dramatic effect on the final 3D product.

For most gradient-plane-based scenarios the furthest point from the observer is termed the vanishing point (VP) [21-23]. The VP may be described as a point in the distant of a scene where straight line edges of artefacts such as buildings, paths and street markings tend to converge (Refer to Fig. 47 (a) and (b)). These inferred imaginary protracted lines are referred to as vanishing lines (VLs). The gradient plane may therefore be described as the gradual gradient variation in an image from the observer to the VP along the direction of the VLs over the range [0, 255] with the furthest point denoted by the value 255. Consequently, the assignment of depth to the pixels, objects and regions is dependent on their location along this so-called gradient-plane [23, 56]. Five of the most common gradient-plane permutations are illustrated in Fig. 54.

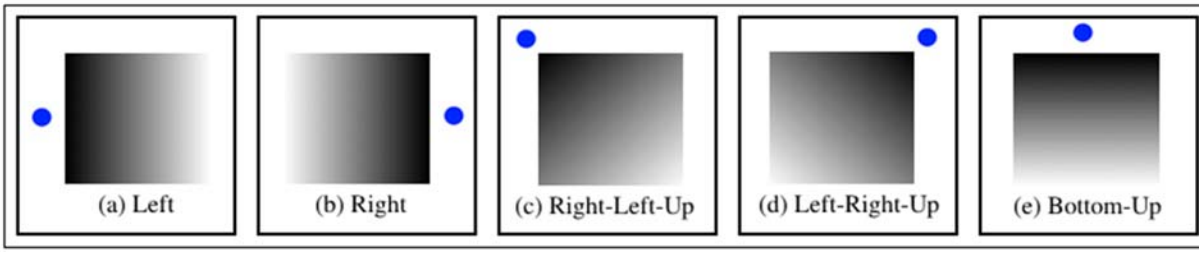


Fig. 54 Five major depth map gradients using the VP (Blue circles denote vanishing points) [136]. The larger the depth value the lower (darker) the grey level intensity.

In this research the method proposed in the section *Robust Vanishing Point Detection of Man-Made Environments* on *p*. 109 is considered for the VP estimation. In the absence of sufficient VLs the VP may be considered as the maximum boundary depth and in the absence of a VP the bottom-up scenario is usually chosen as the default mode [136].

The incremental relative depth values of the ground region may be directly extrapolated from the gradient-plane. Since the sky region denotes the furthest region in the image, every pixel in this region is subsequently assigned a depth value of 255. In Fig. 54 the range of the gradient-planes is [255, 0] whereby the nearest point is denoted by 255 and the furthest point by 0. However, in this research the range is reversed to [0, 255]. The objects on top of the ground region are treated as individual objects and are given slightly shallower (-10) depth prominence compared to their adjacent neighbour pixels. For the objects within the sky region, the relative depths are assigned using a Gaussian re-blurred saliency map. This is discussed earlier on *p*. 94. To account for blur inconsistencies, any value less than 240 is set to 240.

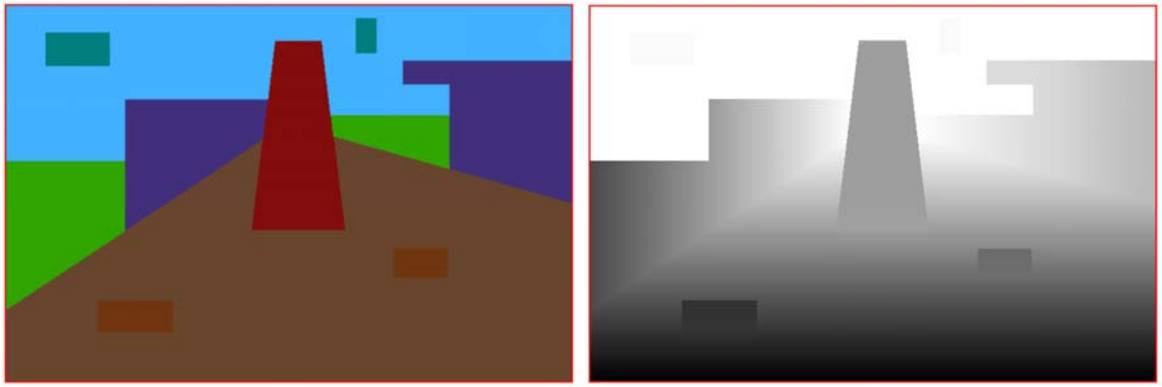


Fig. 55 Relative depth assignment. (a) Gestalt regions. Brown indicates the ground region, green and purple indicates the peripheral ground objects, the OOI is displayed in maroon and the sky region is indicated in blue; (b) Associated depth map. This shows the depth values of the peripheral ground objects following the contours of the ground. The OOI represents an object on top of the ground; therefore, the entire OOI region is assigned the same relative depth value.

The regions on the periphery of the ground region may be broadly described according to two scenarios. There first is where these regions may be considered as objects lying directly on top of the ground region, such as the *Kid*, *Teddy* and *Pumpkin* images in Fig. 53 on *p*. 99. The second is where these regions are located a certain relative distance away from the ground region, such as the *Girl-City* image in Fig. 53. The former is referred to as a continuous scenario and the latter is referred to as a discontinuous or layered scenario.

For type 1, 2, 3 and 6 classifications the ground plane region associated with the ROI, denoted by Π_{GRND}^{INIT}, may be considered as a reference for discerning between the continuous and discontinuous scenarios. A detailed description of this region is discussed on *p*. 61.

In the continuous scenario, the ground region appears to gradually expand across a region of the image through the upward and outward protraction of Π_{GRND}^{INIT} and the pixels adjacent to Π_{GRND}^{INIT} are expected to be similar in intensity to the perimeter of the ground region. In this case Π_{GRND}^{INIT} is considered to be a small subcomponent of the ground region.

In the continuous scenario, depth is assigned separately to each individual column of pixels in each connected component representing the objects on the periphery of the ground region by vertically protracting the depth value of the adjacent ground plane pixel throughout the column. If a column within the object has no associated adjacent ground plane pixel, then the depth of the column is equal to nearest valid column depth value. This proposed technique will allow the depth of the object to follow the contours of the ground; this concept is illustrated in Fig. 55.

In the discontinuous scenario, the ground region is expected to be almost equivalent to Π_{GRND}^{INIT} and the ground region will appear to have an abrupt delineation from the rest of defocussed BG regions. From experiment, if there is a greater than 70% overlap between the ground region and

the straight line bounded representation of Π_{GRND}^{INIT} (described on *p*. 61), then it is considered to be discontinuous. For this eventuality the relative depth values are assigned to regions or objects on the periphery of the ground using a Gaussian re-blurred saliency map. This is discussed earlier on *p*. 94. In this case the relative mean defocus distance between each of the objects on the periphery of the ground and the ground region is considered.

Although the continuous and discontinuous scenarios may apply to all gradient-plane-based images, the likelihood of discerning between them for type 4 and 5 classifications is extremely low. This is owing to the absence of a focussed ground region, as in the *Coke* and *Dog* images in Fig. 53. As a consequence, for these two classifications it is always assumed to be continuous.

For type 1 and 6 classifications the entire depth map is merged with the Gaussian re-blurred saliency map by using the shallower depth value as the dominant intensity. This is to account for some of object discontinuities that may arise during the segmentation process.

iii. Depth Map Finalisation

The techniques employed for the allocation of depth to the objects and regions within the image may result in certain areas exhibiting spurious edges, as well as striated or blocky artefacts. To minimise some of these irregularities, including discontinuities and outliers, a domain transform edge-preserving filter is applied to the depth map. The filter is discussed earlier on *p*. 95. In the proposed model, from experiment, the spatial standard deviation is set to 20, the range standard deviation is set to 0.1 and the number of iterations is set to 3. To account for the full grayscale spectrum, the final step involves the extrapolation and distribution of the smoothed pixel intensities across the range [0, 255]. This is illustrated in Fig. 50(c) and Fig. 53(c). This concludes the method proposed for the unsupervised generation of a depth map from a single 2D LDOF image.

D. Results and Discussion

The training and test data is sourced from previous works as well as numerous images from the World Wide Web (WWW). Comparative analyses are performed against six other proposed methods [60-65]. Results of some of the benchmarking and test images obtained from the WWW are also presented.

Some of the representative results obtained for the benchmarking data are illustrated in Fig. 56. The original images in Fig. 56(a) are titled (from top to bottom) as *Kid*, *Pumpkin* and *Bird-Reed*, respectively. Fig. 56(b) show the results obtained from the methods proposed by Zhuo and Sim [63]. Fig. 56(c) present the results of the proposed method.

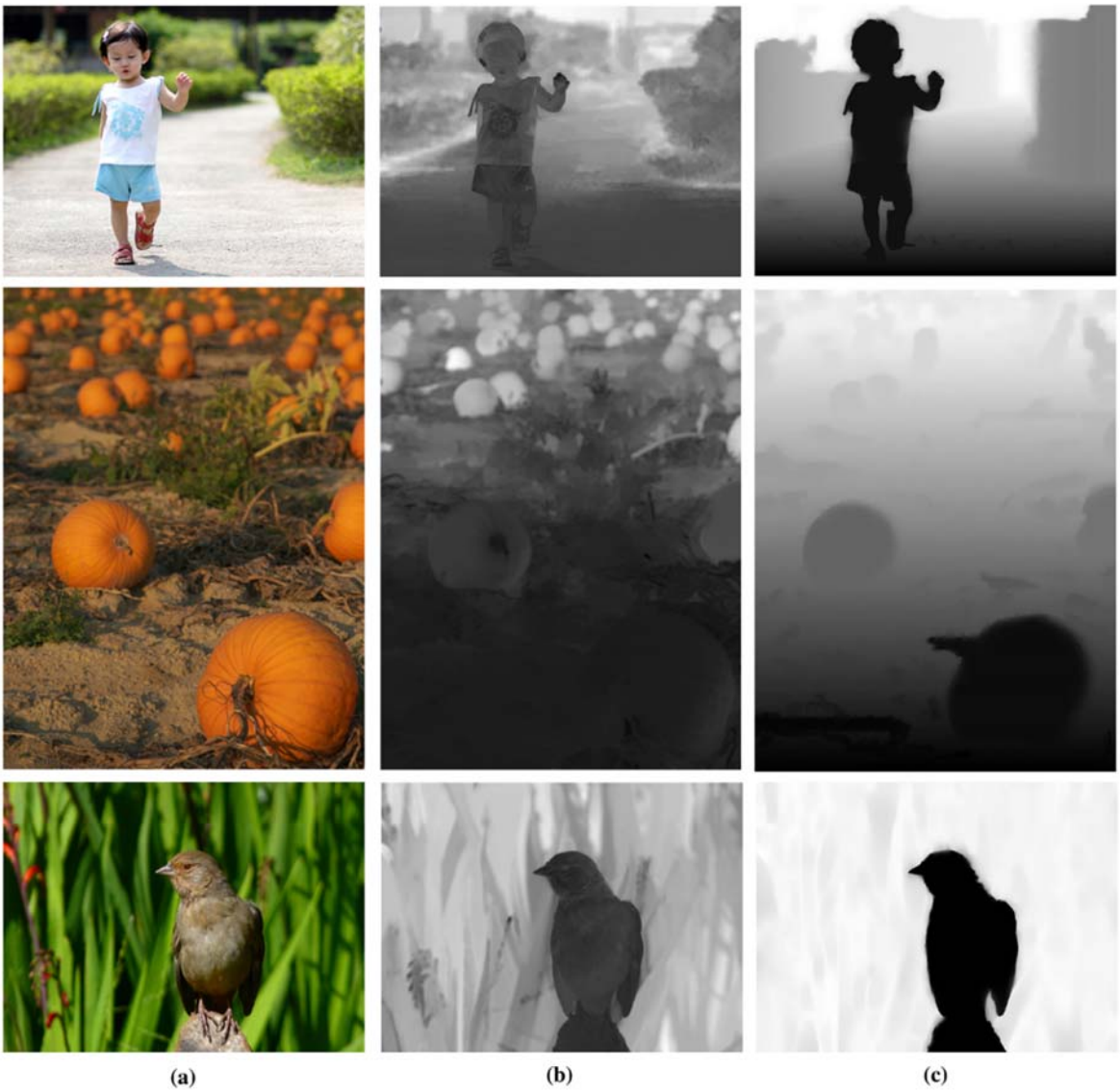


Fig. 56 Benchmarking: *Kid*, *Pumpkin* and *Bird-Reed* (from top to bottom). (a) Original image; (b) Defocus map generated by Zhou and Sim [63]; (c) Depth map generated using proposed method.

When considering Zhuo and Sim's method, although the three defocus-based depth maps appear on face value to be aesthetically more appealing, they nevertheless in practical terms contain numerous inconsistencies. In the *Kid* and *Bird-Reed* images it is evident that the boy and bird, respectively, represent single closed boundary objects in the images. However, in their interpretation these objects incorrectly exhibit varying degrees of relative depth. Moreover, in the *Kid* and *Pumpkin* images there are numerous objects and regions on top of the ground plane region that have been allocated vastly inconsistent relative depth values.

In the proposed model the OOI is considered to be a single object that is equidistant from the camera and therefore the entire closed connected region is allocated the same relative depth value.

Moreover, the objects on the ground region, such as the kid (OOI) and hedges in the *Kid* image and the pumpkins in the *Pumpkin* image, are allocated relative depth values that are more consistent with their observed 3D Gestalt placement in the scenes.

Although there exist some minor relative depth inconsistencies in the proposed method, it nevertheless shows a noticeable improvement to Zhuo and Sim's approach.

Fig. 57 Subjective comparison 1. (a) Original image; (b) Defocus magnification blur map generated by Bae and Durand [60]; (c) Defocus map generated by Zhou and Sim [63]; (d) Blur radius map generated by Zhu et. al [64]; (e) Depth map generated using proposed method.

Some of the representative results obtained for the test data are illustrated in Fig. 57. The original image is illustrated in Fig. 57(a). Fig. 57 (b) – (d) show the results obtained from the methods proposed by Bae and Durand [60], Zhou and Sim [63] and Zhu et. al [64], respectively. Fig. 57(e) present the result of the proposed method.

The same concerns highlighted in the analysis of Fig. 56 above are applicable to the results generated by the other approaches in Fig. 57. The proposed model shows a noticeable improvement to Bae and Durand's method and is comparable to the method proposed by Zhu et. al.

A second set of comparative test results is illustrated in Fig. 58. The original image is illustrated in Fig. 58(a). Fig. 58(b) and (c) show the results obtained from the methods proposed by Valencia et. al [62] and Guo et. al [61], respectively. Fig. 58(d) present the result of the proposed method.

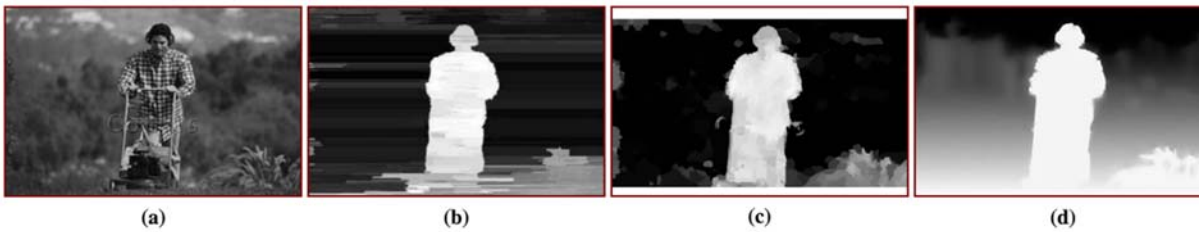


Fig. 58 Subjective comparison 2. (a) Original image; (b) Depth map generated by Valencia et. al [62]; (c) Depth map (cropped in the reference paper) generated by Guo et. al [61]; (d) Depth map generated using proposed method.

When considering the method proposed by Valencia et. al, although the segmentation of the foreground objects as well the subsequent assignment of relative depths to these regions may be considered as reasonably accurate, essentially the all the background regions are negated. In comparison to this approach, while the method proposed by Guo et. al. shows more refinement of the boundaries of the objects and less noise and stripy artefacts, it also disregards the background regions.

Although there exist some minor relative depth inconsistencies in the proposed method, it nevertheless deals more accurately with the foreground objects and more effectively with the assignment of relative depths to the background objects and regions.

A third set of comparative test results is illustrated in Fig. 59. The result obtained from the method proposed by Ko et. al [65] is shown in Fig. 59(b) and the result of the proposed method is provided in Fig. 59(c).

When considering the method proposed by Ko et. al, although the segmentation of the foreground object may be considered as reasonably accurate, a significant amount of the background objects and regions are assigned inconsistent relative depth values and in particular, the ground plane region, which should exhibit incremental relative depth towards the vanishing point or region.

Fig. 59 Subjective comparison 2. (a) Original image; (b) Depth map generated by Ko et. al [65]; (c) Depth map generated using proposed method.

Although there exist some minor relative depth inconsistencies in the proposed method, it nevertheless deals more accurately and effectively with both the segmentation and the subsequent assignment of relative depths to the background objects and regions.

Some representative results of the VP estimation are illustrated in Fig. 60. A detailed description of this sub-component of the depth map process as well as analysis of the results are presented in the section *Robust Vanishing Point Detection of Man-made Environments* on *p*. 109.

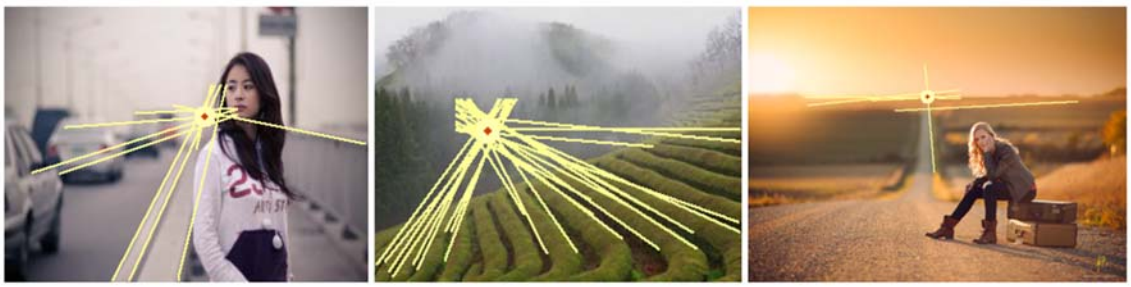


Fig. 60 Vanishing point estimation.

The testing computer is a Quad Core CPU 2.50 GHz with 4.00 GB of RAM. The proposed unsupervised method is implemented using MATLAB R2014a. For equidistant-based images the model achieves an average processing time of 1 s for an average image size of 543×385. For gradient-plane-based images the model achieves an average processing time of 3.3 s for average image size of 596×420.

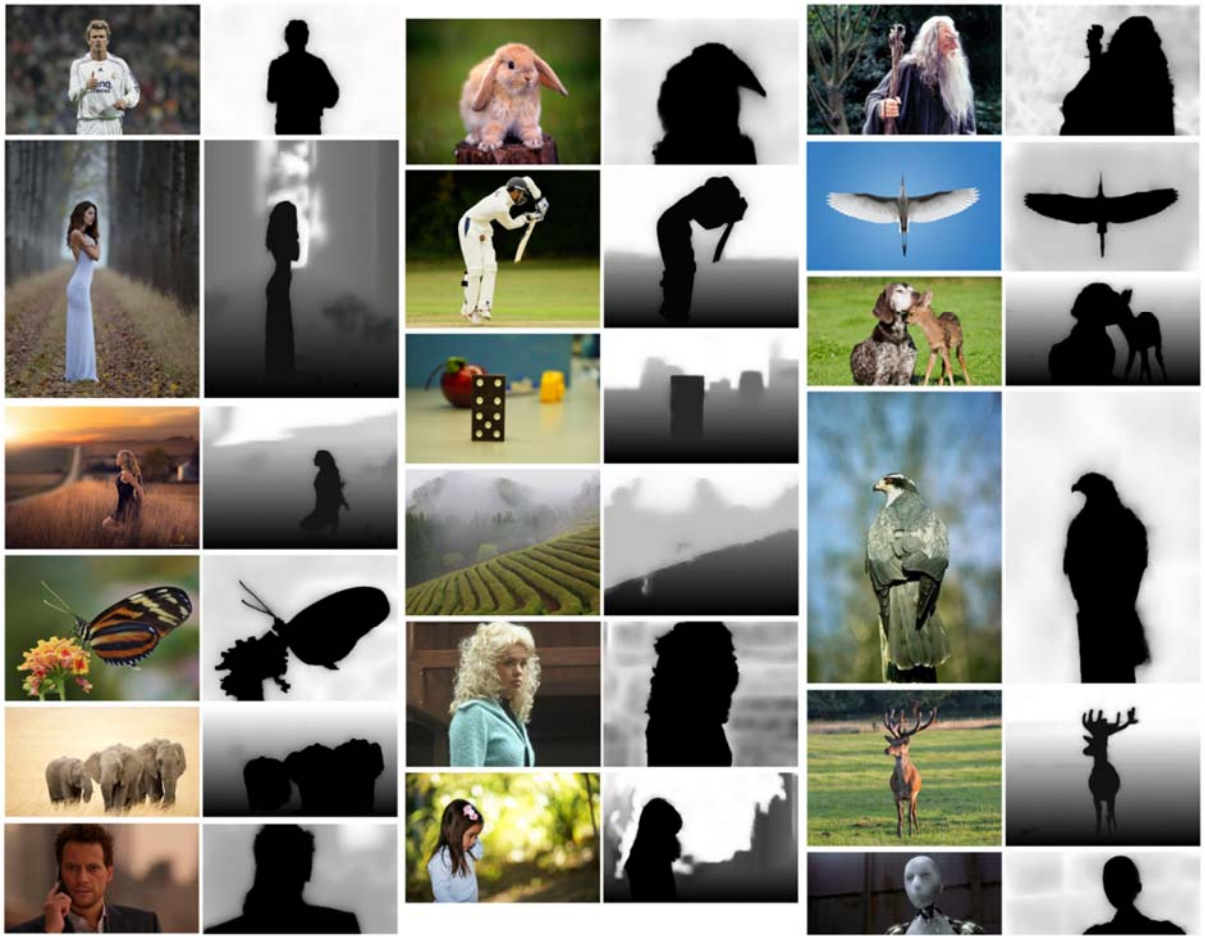


Fig. 61 Depth map results.

Owing to space constraints only a selected number of the depth map results are presented in this section. The author may be contacted at serenr@gmail.com for more than 300 LDOF images and their extrapolated depth maps.

Currently in the literature there is insufficient objective ground-truth depth map data available for LDOF images. As a consequence the presented results are only subjectively assessed. Future work may include the generation of a ground-truth depth map database for LDOF images. Although the generation itself will be subjectively produced it may be objectively assessed through peer consensus.

Even though an objective comparison is not provided it must be emphasised that the relative depth maps autonomously produced by the proposed model are not meant to be precise depth representations of all the delineated objects and regions in the image; but rather as sufficient approximations of the relative depths that may be applied specifically for use in the extrapolation of stereoscopic image pairs from single LDOF images using depth image-based rendering.

A limitation of the proposed approach is that it is inherently dependent on the accuracy of the segmentation. However, the results show that by applying Gestalt principles to the segmentation together with vanishing point analysis and Gaussian blur saliencies it is possible to overcome the ill-posed problem and produce a reasonably acceptable relative depth map of a LDOF image.

Another limitation is the assumption that the entire OOI is at a uniform relative depth. For close-up images this assumption may conflict with the sensitive a priori visual expectations of an OOI, such as the changing contours of a human face. Although the relative depth map may be clean and well defined it may nevertheless result in the final 3D stereoscopic image exhibiting a cardboard effect. Further research, possibly using machine learning approaches, is warranted in improving this shortcoming.

E. Conclusion

A novel unsupervised method is proposed for the assignment of relative depths to the objects and regions within a single 2D LDOF image. Depending on the description of the non-OOI region, relative depths are assigned using either Gestalt principles together with vanishing point detection or Gaussian blur analysis. In the proposed model there are two descriptions of the non-OOI region in a LDOF image. These include equidistant-based and gradient-plane-based. The former only applies to type 4 and 5 classifications while the latter may apply to all classifications. A novel method is proposed for the discernment of these two scenarios through the analysis of the Gestalt-based ground and sky regions.

For the equidistant-based scenario the entire non-OOI region may be described as being a certain relative depth behind the OOI. The results show that for these scenarios Gaussian re-blur analysis may be effectively employed for the assigning of relative depths across the image.

For the gradient-plane-based scenario the non-OOI region is constituted of distinct objects and sub-regions. This may include a ground region as well as objects, such as the OOI, that are located on top of this plane at different relative depths. The results show that each object or sub-region within the non-OOI region may be effectively delineated through the correlation of *k*-means and binary quantisation segmentation and subsequently labelled into one of five proposed Gestalt-based categories. Moreover, the results show that a gradient-plane may be accurately constructed by correlating the Gestalt-based ground region with vanishing point detection and subsequently employed in the assignment of relative depths to the objects and regions across the image.

Results also show that blocky artefacts and inconsistencies in the depth maps may be adequately alleviated through the application of a domain transform edge-preserving filter.

Future work may include the use of more scenario-specific segmentation methods as well as the possible incorporation of machine learning to more effectively identify and subsequently assign relative depth values to the objects and regions in an image according to the specific semantics of the involved scene. Moreover, the proposed method may be expanded to account for high DOF images.

V. VANISHING POINT DETECTION OF MAN-MADE ENVIRONMENTS

A. Introduction

Vanishing point (VP) detection plays a vital role in 3D interpretation as well as in arenas such as machine vision, depth estimation as well as 2D-to-3D conversion systems. VP estimation is a subcomponent of linear perspective (LP). LP is the term used to describe one the monoscopic depth cues employed by the human visual system to explain the phenomenon of 3D depth perception on a 2D plane through the inference of imaginary converging lines or regions. The more the number of lines or regions converge, the more it will have the appearance of being further away [21-23]. The point or apex at which these imaginary *vanishing* lines (VLs) or regions (VRs) terminate is termed the vanishing point (VP). This is illustrated in Fig. 62.

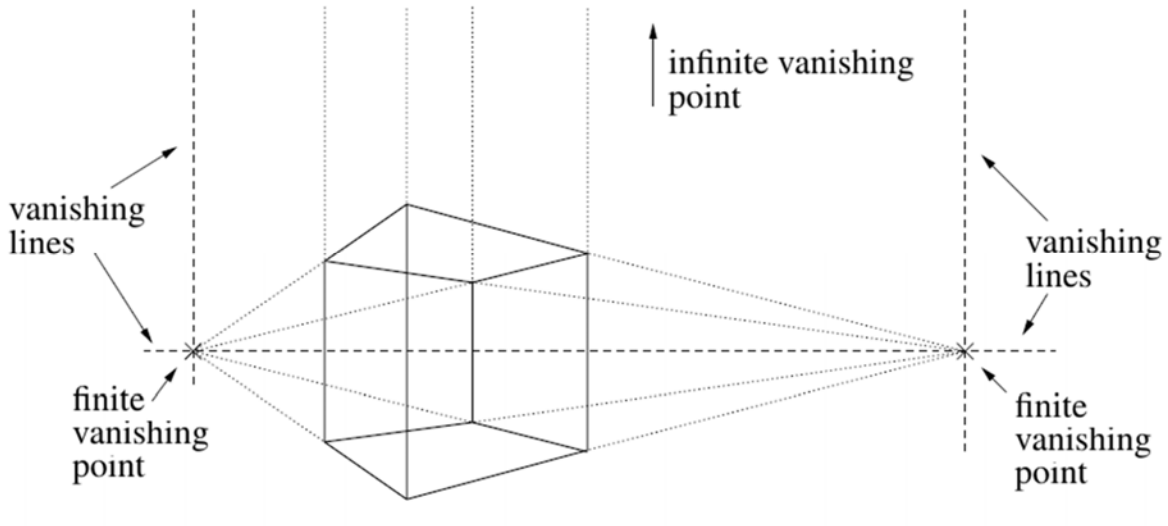


Fig. 62 Vanishing lines and points [137]

Features, like these VLs, VRs and VPs as well as other useful topographical data may be identified and extrapolated using several geometric-based algorithms. Natural settings are more prone to irregularities compared to the regular structures present in artificial (man-made) environments where there is a higher likelihood for the existence of straight as well as parallel lines [138]. In the former scenario clues may possibly be extracted through analysis of clustered textured regions. For the latter situation clues may be extracted based on the horizon, as well as straight lines of objects in image frames, which among others may include buildings, walls, roads, street markings, piers and railway tracks.

The precise location or interpretation of the VP is often subjective and there is no optimal solution for the unsupervised extraction of these phenomena. Although there have been numerous

algorithms proposed for the estimation of VPs across a variety of structural scenarios, the problem nevertheless remains ill-posed and most of the solutions, more often than not, require complicated computational modelling and/or significant a priori knowledge about the object space, despite their claimed practicability [57].

This research proposes a novel and simple technique for the estimation of the VP and associated VLs based on a preconceived notion of converging straight lines in images of man-made environments through the analysis and extrapolation of data entirely in the image plane coordinate space. This research may be considered as an amalgamation and extension of the general techniques proposed by Schaffalitzky and Zisserman [59], Almansa et. al [139], Battiato et. al [23] and Tsai et. al [140].

The supervised training of the algorithm is performed using images from previous works [23, 56-59] as well as numerous random images from the World Wide Web.

The next section of this paper discusses some of the approaches that have been previously proposed for straight line detection and VP estimation. The proposed method is presented in section C. A block-based approach is proposed for the extrapolation of the candidate VPs from an image space-based accumulator array. An evaluation metric is also proposed for the selection of the optimal VP and the associated VLs. Section D provides a demonstration and discussion on the performance of the proposed approach on real image data. Finally, conclusions and recommendations for future research are stated in section E.

B. Previous Approaches

There currently exist three broad types of approaches to VP estimation. These include edge-based [137-140], region-based [141-143] and prior-based [144-146] methods.

VP detection using edge-based methods is principally based on the analysis of straight line edge segments in an image. These methods fundamentally require three processes [137]. The first process, which is usually performed subsequent to edge detection, involves the interrogation and *identification* of the straight lines in an image. The second process, which is referred to as the *accumulation* step, comprises the clustering of the dominant straight line segments according to common candidate VPs. The third and final process, which is referred to as the *search* step, involves the estimation of the dominant VP and associated VLs.

The precise detection of edges is key to the identification of the dominant lines in an image. The most common edge detection techniques usually involve the iterative application of isotropic masks to every pixel location across an entire image. The edges are subsequently extrapolated through thresholding of the estimated horizontal, vertical and, in some instances, diagonal gradients of each pixel. There are several methods employed for edge detection, like for example

the popular Canny algorithm [95]. Other gradient-based masks, such as Frei-Chen [138] and Sobel [23, 57], have also been considered for use in VP detection.

The efficacy of an edge map is dependent on the chosen threshold value, which is primarily dependent on intensity and not actually on the relationships between pixels. If the threshold value is too low, then both false as well as thick edges may result. On the other hand, if the threshold value is too high, then some edges may be too discontinuous or are not detected at all. Two popular thresholding techniques are the maximum entropy and Otsu methods [147].

Owing to several factors, like for example lighting, continuous pixel intensities and blurs, it is plausible to expect that some of the edges will either not be detected or will be discontinuous in their representations of objects and regions. As a consequence it is common practice to refine the edge map [94]. Noise may be reduced through, for example, the removal of small connected components as well as the pruning of spurious edges. Disconnected boundaries may be linked through, for example, the use of morphological reconstruction methods. For the linking of close neighbourhood regions simple morphological methods, such as dilation and erosion, may be employed. However, to account for several scenarios, which may include larger discontinuities, a more adaptive morphological edge-linking algorithm may be required [148].

Edges in general are too broad in their description to be effective for VP detection. For this purpose, *straight* edges or more specifically, *dominant* straight lines, are more appropriately suited. One popular method of straight line detection is the Hough Transform (HT) technique [149]. It is often referred to as the Hough/Radon transform owing to its mathematical equivalence to the Radon transform [150]. In this case the dominant straight lines in an image are extrapolated through analysis of the slope or tangent and the y-axis intercept of collinear pixels in the binarised edge map. Initially, each parameter pair are sampled at suitable locations and collated into an accumulator matrix, also referred to as an *accumulator space* or *array*. Subsequently, through examination of the maxima values, the dominant straight lines may be identified. The key benefits of the HT include its tolerance to gaps in boundary locations and being largely immune to noise.

An alternative technique to the HT straight or dominant line detection method involves the analysis and identification of the straight lines directly in the image plane coordinate space. In the approach proposed by Schaffalitzky and Zisserman [59] the straight lines are extracted directly from the binarised edge map through the breaking of edges at points of high curvature.

Another approach involves the application of the Helmholtz principle [139, 151, 152]. This technique involves the detection of similar groups or types of lines, like for example level or straight lines, through the computing of thresholds in images containing geometric structures, such as those associated with artificial environments. This method is not considered as another type of global edge detection technique but rather as a means of predicting potential candidate edges based on some feasibility criteria rather than on a priori edge information.

A third technique proposed for straight line segment detection in the image plane is the block-based approach [140]. In this case each pixel in the middle position of an edge segment is analysed and adjacent pixels are plotted in an outward direction away from the respective central reference pixel until the adjacency rule becomes invalid in both directions; the resulting sub-segment then represents a straight line in the image, as illustrated in Fig. 63.

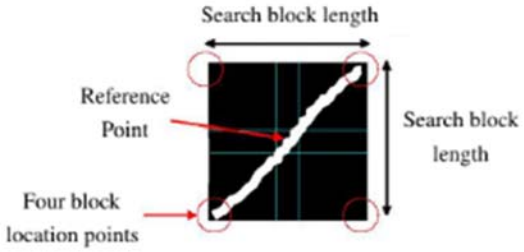


Fig. 63 Block object finding [140]

The detection or identification of straight lines in an image is an important first step. However, in order for the straight line segments to be useful in VP detection it is necessary to be able to analyse and evaluate the intersection of these lines. In addition to the detection of straight lines, for current purposes, the HT accumulator array may also more importantly be applied to the detection of inter-connected lines [94, 153]. By partitioning the aforementioned accumulator spaces into sectors or so-called *accumulator cells* they may be used to identify potential candidate VPs based on the accumulation or clustering of overlapping prominent lines or their mathematical equivalent.

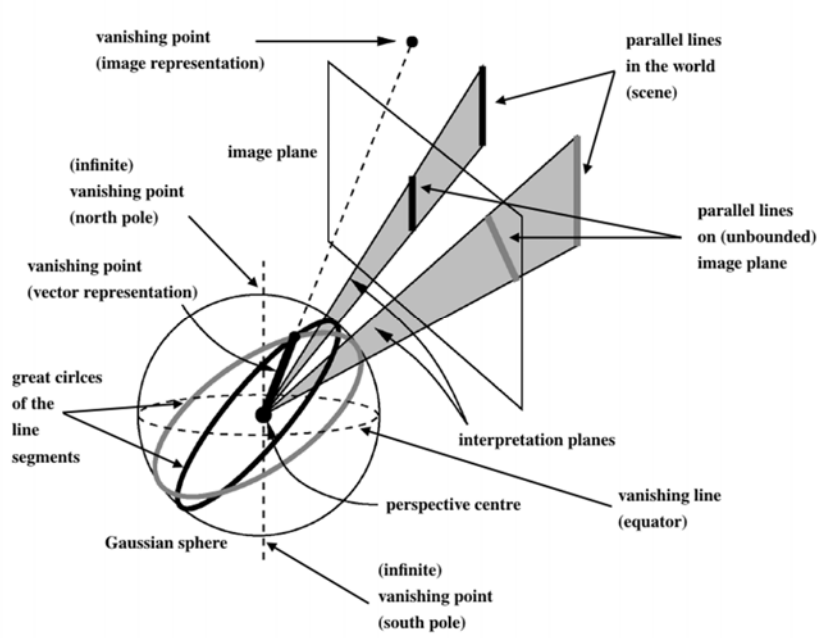


Fig. 64 Gaussian-sphere as an accumulator space [154].

Depending on the angular attributes of some of the prominent lines, there may be scenarios where the candidate VPs are located outside of the image plane or seemingly infinite in nature. To account for this, it may be necessary to translate the unbounded image plane coordinate space onto a bounded space. Two proposed approaches include the Gaussian-sphere [139, 154, 155] and polar-parameter space [138, 155] techniques.

The Gaussian-sphere accumulator array approach was originally proposed in the seminal work by Barnard (1983) [156]. This method employs the use of a Gaussian sphere as a means of representing the finite but unbounded image plane as a bounded space. As illustrated in Fig. 64, each line segment on the image plane, in relation to the perspective centre, is initially transposed or warped onto a Gaussian sphere and then geometrically rendered into great circles. Subsequently, the VPs are hypothesised through the probabilistic interrogation of the largest cluster regions of the overlapping great circles on the accumulation array. Finally, based on the orientation of the associated accumulator cells, the VPs are subsequently extrapolated back onto the image plane.

Since Barnard there have been several refinements and advancements to this proposed technique [157, 158]. However, further interrogation in more recent works show that camera calibration parameters, viewing angle and distance to the scene may create inconsistencies when estimating maxima in the accumulator array. Furthermore, textural effects of natural images may be responsible for sporadic variations and, in combination with artefacts of image geometry, may produce spurious VPs on the Gaussian sphere [137, 138, 154].

As an alternative to the bounded Gaussian-sphere space approach, the prominent lines detected in the unbounded image space may be translated and analysed in the bounded polar-parameter space. These polar-parameter space approaches are inherently based on the application of the HT. Tuytelaars et al. (1998) [155] proposed the use of a cascaded HT (CHT) method whereby the HT is iterated through three stages in the polar-parameter space. As a means of preserving the symmetry between the spatial domain and the polar-parameter space, the original unbounded space is initially split into three bounded subspaces and the CHT is applied separately to each subspace. The first stage of the CHT involves the application of the HT to the original image, which in turn yields the peaks corresponding to the straight lines in the image. The second stage of the CHT involves the application of the HT to the filtered output of the first stage, which in turn yields peaks corresponding to intersecting lines in the image. The maxima of the dominant peaks will typically correspond to the VPs. The third and final stage of the CHT involves the application of the HT to the refined output of the second stage, which in turn yields peaks corresponding to *dominant* collinear line intersections in the image. These dominant lines may often represent significant boundaries of interest, such as the horizon line. This aforementioned proposal was originally developed for the interrogation of aerial images. However, adaptations to

the method for application to more common scenarios, like landscapes, roads, buildings and indoor structures, have been proposed [23, 57, 138].

As an alternative to the Gaussian-sphere and polar-space approaches, an image-plane coordinate space approach is also proposed. For the former two techniques the dominant straight lines are effectively translated from an unbounded image space into an alternative bounded space for estimation of the VP. However, for the latter method, it is proposed that the respective prominent straight lines be directly analysed in the unbounded image-plane or spatial domain [138, 140, 159].

VP and VL estimation in the spatial domain involves the direct computation of intersecting dominant straight lines in the coordinate space. This technique involves four processes. Firstly, converging or concurrent dominant lines are generated by effectively protracting each dominant straight line across the entire image window. Secondly, candidate VPs are selected as points (or clusters) in the image with the most intersections around them. Thirdly, the dominant VP is chosen from the collection of candidate VPs. The second and third processes are essentially equivalent to the accumulator array method of the HT. Fourth and finally, the VLs are chosen as the extended prominent straight lines passing through or close to the VP [23].

The spatial domain-based approaches attempt to eliminate some of the false alarms and inconsistencies, like for example asymmetry, observed in the Gaussian-sphere- and polar-plane-based approaches. However, the concern with the former is that the accumulator space is essentially unbounded. As a consequence, three concerns need to be addressed when analysing the dominant straight lines in the image plane. The first involves the instances where the VP may lie beyond the image window. For these situations it becomes necessary to expand the parameter space so as to accommodate those intersecting points whose coordinates are greater than the image size. The second involves the handling of possible infinities that may arise owing to non-converging vertical, horizontal and parallel lines. The third involves the situation whereby more than one VP needs to be taken into account.

Most spatial domain VP detection techniques employ a sort of joint Gestalt [160] approach when analysing the straight lines in the image. Some of these Gestalts include the orthogonality between lines and their subsequent convergence in 3D space [139, 140, 154] as well as parallel, coplanar and equally spaced lines [59].

In terms of orthogonality between lines, one option is to consider an angular precision factor [139, 154]. Another may be to employ threshold constraints, such as those in illustrated in Fig. 65. In this case the constraints include the angle θ and distance d between two regression lines, denoted by L_1 and L_2 [140]. This is mathematically expressed as

$$\tan\theta \leq \tan\theta_{threshold}, \tag{37}$$

where

$$\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right| \text{ for } m_1 m_2 \neq -1, \tag{38}$$

and

$$d \leq d_{threshold}. \tag{39}$$

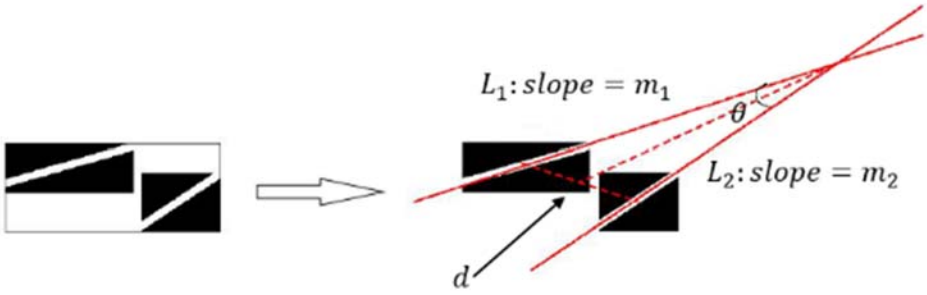


Fig. 65 Object combination scheme for two regression lines [140]

Both Schaffalitzky and Zisserman [59] and Almansa et. al [139] show that for images of artificial environments the analysis of the 3D parallelism alone is a reliable method for estimating VPs, including those that lie outside the boundaries of the image window.

In the case of the Schaffalitzky and Zisserman approach the concurrent straight line segments are firstly grouped using the random sample consensus (RANSAC) technique [161]. Secondly, the coplanar equally spaced parallel lines are extracted by performing a direct analysis of the binarised edge map as well as the analysis of additional parallel correspondences and patterns through the interrogation of feature points in the image. Thirdly, a set of probable candidate VPs are estimated using a maximum likelihood estimate (MLE) threshold. The MLE is based on the perpendicular distances between the estimated crossing points and the associated converging lines within a localised region. Fourthly, the coplanar parallel lines are translated into a single projected vanishing line. Finally, the dominant VP and associated VLs are extrapolated through the correlation of the projected vanishing line with the candidate VPs.

In the Almansa et. al approach the straight line segments are initially grouped according to rotation and translation and subsequently projected onto a vanishing region. Each vanishing region may be described as a quadrilateral containing a group of concurrent lines. The grouping of the lines are based on a practical application of the Helmholtz principle [162] whereby each quadrilateral is firstly, positioned according to the equiprobability that the lines meet a specific vanishing region and secondly, sized and shaped according to the angular precision the respective lines. The dominant vanishing region is estimated through the local neighbourhood analysis of the

clustered maxima of the refined vanishing regions together with the thresholding of the lines based on a minimum description length.

Edge-based approaches provide an effective method of VP estimation without the need for camera calibration or a priori scene information. In addition, these methods may be able to account for finite as well as infinite VPs.

Region or characteristic-based approaches to VP detection are based on the exploitation of characteristics, such as texture and colour, measured around a local neighbourhood within an image. By analysing the dominant orientation of texture within a region in an image using, for example, Gabor filtering [163], dominant orientation rays may be identified and, based on their intersection, used to detect the VP.

Another region-based approach is to consider the use of colour segmentation. In these cases, the image is initially delineated into segments, using for example the mean shift technique [23, 56, 57]. Subsequently, the segments are grouped and assigned to a set of pre-defined regions, such as the foreground, skyline, road and mountain. Finally, the VP is estimated through analysis of the regions. This approach is illustrated in Fig. 66.

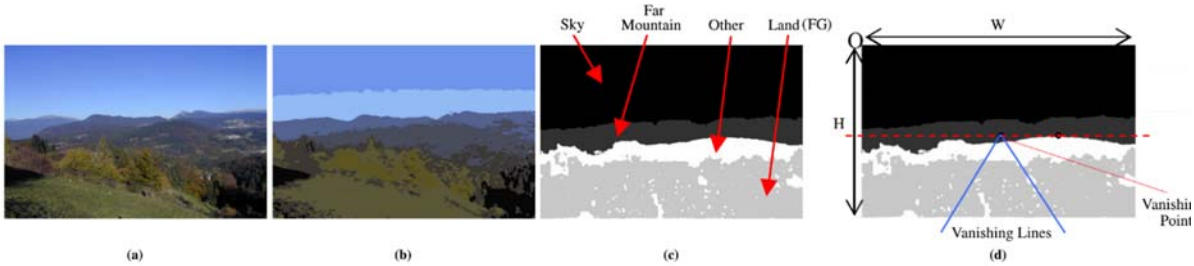


Fig. 66 Region-based VP detection [56]. (a) Original image; (b) Segmented image; (c) VP based on image classification; (d) VP and VL detection using an edge-based method.

Region-based methods are often dependent on the relationship between the positions and characteristics of objects in an image. This type of grouping is termed the Gestalt laws of perception [24]. For example, it may be assumed, if a sky region and accompanying horizon line exist in an image, it will have a higher probability of being located towards the upper part of the image and vice-versa for the ground region. Furthermore, the smaller the objects appear to be, as well as the closer the objects are clustered together, may provide an indication of where the VP is possibly located.

Region-based approaches are more effective than edge-based techniques with regards to images of natural scenes. However, for artificial environments the latter approaches have shown to be more effective. For a more globally robust method of VP detection it may be advisable to consider a combination of edge- and region-based techniques.

An attempt to combine edge- and region-based methods has been proposed by Battiato et. al [23]. In their approach images are initially classified into different types of scenarios, such as landscape as well as outdoor and indoor man-made environments. The edge-based component of their method is employed in the case of the latter two scenarios and involves the use of the HT accumulator technique for the detection of both the main straight lines as well as the intersections of the main straight lines. The region-based component is employed for use in landscape type scenarios. This technique involves the initial delineation of the image using mean shift segmentation. Subsequently, each cluster is labelled as a distinct semantic region like for example, ground or sky. Finally, the VP is extrapolated through a sort of rudimentary analysis of these sub-regions. This approach is illustrated in Fig. 66.

Another combined method has been proposed by Chappero et. al [57]. Their approach may be considered as an extension of the method proposed by Battiato et. al, since both techniques involve the use of the HT accumulator techniques and both rely on the analysis of colour clusters that are delineated through the application of mean shift segmentation. However, in the Chappero et. al approach an added criterion involving the analysis of pixel gradients is considered for the identification of the candidate VLs.

Other approaches to VP detection involve the combination of edge- or region-based methods together with prior-based techniques [144-146]. However, these approaches are not robust in determining VPs and subsequent VLs since the estimated results are mainly dependent on large iterative training sets, adjustments to numerous variables and modifications to several reasoning procedures. In addition, spurious results may occur owing to scarcity of object identification as well as sparse and vague object relations and distributions in the image. Under very narrowly defined conditions, prior-based methods have shown to perform slightly better than the non-prior-based approaches. However, in these cases accuracy still has to be weighed against the increased and costly processing time.

C. Proposed Method

This research provides a simple and robust method for the detection of VPs and associated VLs in images of artificial or man-made environments. This is achieved through the direct extrapolation and analysis of the converging straight lines entirely within the image plane coordinate space. The following process provides a high-level overview of the proposed method:

Step 1: Perform edge detection and morphological refinements
Step 2: Identify dominant or major straight lines
Step 3: Refine the straight-line segments

Step 4: Construct an accumulation matrix based on the extrapolation of intersecting points of every straight-line segment with every other straight line segment through solving of the polynomial from their respective slope-intercept forms.

Step 5: Construct a second accumulation matrix based on the perpendicular distance of every straight line to every intercept point in the first accumulation matrix.

Step 6: Select up to 5 candidate VPs from the second accumulation matrix by considering the number of segments (associated VLs) passing near each intercept point within a certain range.

Step 7: Select the optimal or dominant VP through a deeper interrogation of the associated VLs passing near each intercept point within a certain range.

1) Straight Edge Segment Detection

Edge detection is performed on the source image using the Canny method. Based on the evaluation of the training data a threshold of 0.25 and a sigma of $\sqrt{6}$ is chosen. If the input image is colour, then the Canny algorithm is applied separately to each R, G and B layer as well as to the grayscale version of the image. The four binary maps are subsequently merged into a single binarised edge map. This is illustrated in Fig. 67(b).

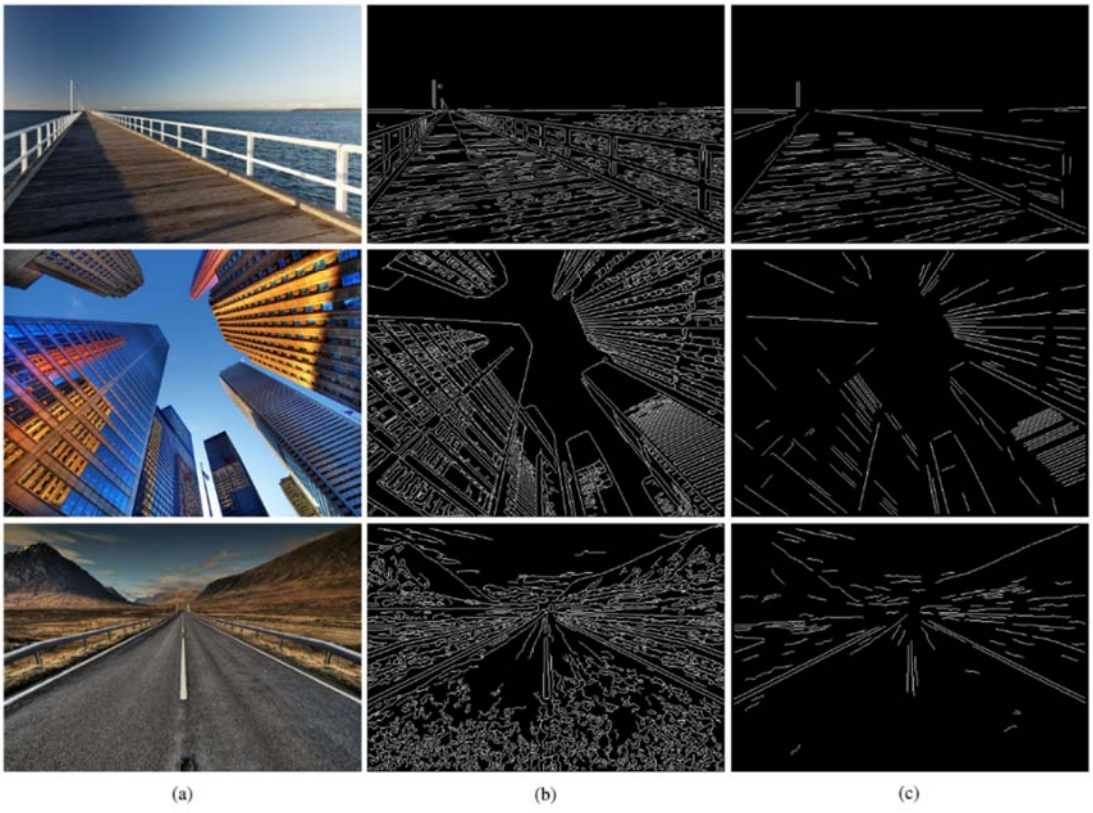


Fig. 67 Detection of straight line edges. (a) Original images; (b) Full edge maps; (c) Straight line segments.

Only three types of quantised straight edges may exist in an image viz. horizontal, vertical and diagonal; of these the diagonal or sloping edges are the most significant in terms of VP detection. Majority of the segments in both the horizontal as well as vertical edge maps will contain segments that are sloping i.e. between 0° and ±90° (excluding ±45°). Edges sloping at ±45° are accounted for in the diagonal edge map.

When considering a single row within an edge map, any absolutely vertical (±90°) or diagonal (±45°) line will only be represented by a single pixel within the row. Based on this understanding, the horizontal edges are then associated with all connected components containing 2 or more pixels within a single row. In the proposed model the horizontal edge map is produced by iterating through each row and extracting any connected component containing 2 or more pixels and collating the segments into a single binarised array.

Conversely, when considering a single column within an edge map, any absolutely horizontal (0°) or diagonal (±45°) line will only be represented by a single pixel within the column. Based on this understanding, the vertical edges are then associated with all connected components containing 2 or more pixels within a single column. In the proposed model the vertical edge map is produced by iterating through each column and extracting any connected component containing 2 or more pixels and collating the segments into a single binarised array.

The diagonal edge map is produced by removing all horizontal and vertical intersections from the original edge map. The remaining segments are expected to be small clusters, single pixels, and some edges containing slopes around ±45°.

For VP estimation it is understood that the dominant straight line edges in an image play a more prominent role. As such, a minimum descriptor length (MDL) threshold is applied to the horizontal, vertical and diagonal edge maps. This results in the removal of all connected components containing pixels fewer than the MDL. From experiment, the MDL is calculated as 0.01% of the pixel area (width multiplied by the height) of the image. The three edge maps are subsequently merged into a single binarised edge segment map. To finalise the edge segment map, the locations of all corner, crossing pixels, spurious edges (length of 3 or less pixels) and closed loop or connected border regions are set to zero, followed by application of the MDL threshold. Rudimentarily, any segment containing branchpoints or more than two endpoints may be considered as being non-straight edges; as such, any edge fitting this description is removed from the edge segment map. The remaining edges in this initial straight edge segmentation map are further interrogated through a more precise measure of straightness.

A simple method of determining if an edge is a straight line may be to calculate the difference between true length (actual number of pixels) and straight line length (quantised pixel distance between endpoints); if the difference is less than or equal to a certain threshold, like for example 3 pixels, then the edge may be considered as a straight line.

Another method may be to consider the area or energy contained within the closed region of each segment. If one were to visualise a line segment, then the joining of the endpoints by a straight line will form either a closed region or a single connected component with multiple closed regions. The minimum width of the quantised area or energy within the closed regions may provide a means of determining whether the line is considered to be straight or curved. In the case of a straight line segment the minimum width of the area is expected to be zero or at the very most infinitesimal. In pixel terms, from experiment, this is denoted by connected components containing minimum thickness of only 1 or 2 pixels. Conversely, a curved (non-straight) segment is expected to be comprised of relatively large energy regions. From experiment this taken to be regions containing minimum widths greater than 2 or more pixels.

In the proposed model the latter technique is chosen for the identification of the straight edge segments. Although this technique is computationally more expensive than the former approach, from experiment it is, nevertheless, shown to be more accurate. Fig. 67(c) provides an illustration of the filtered straight line edges in the image.

2) Straight Line Interpretation and Co-Planar Refinements

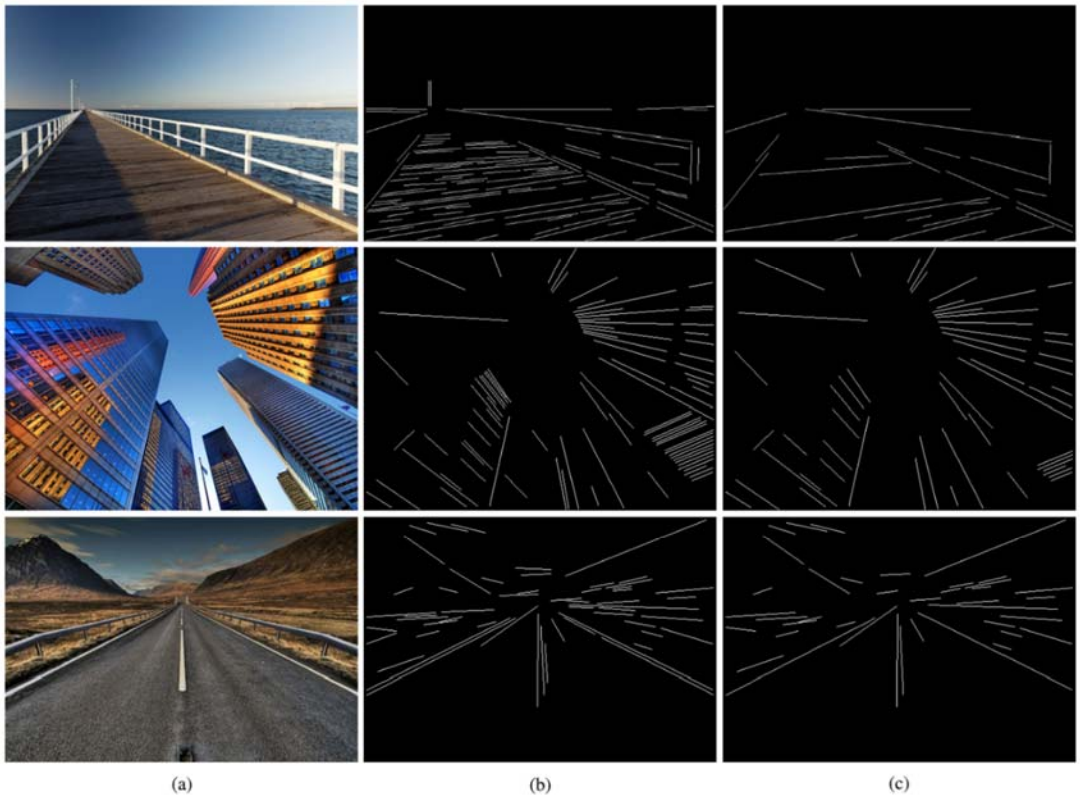


Fig. 68 Digital straight lines. (a) Original images; (b) Digital straightened segments; (c) Refined digital straight lines.

Majority of the valid binarised edges only provide an approximation of the straight lines or edges in the image, however, for VP estimation these edges need to be perceived as being absolutely (in relative terms) straight. To resolve this, the endpoints of each segment are considered, since only 2 points are required to completely describe and represent a straight line in 2D coordinate space. An illustrative interpretation of the straight lines is shown in Fig. 68(b).

To minimise infinities only the dominant (longest) horizontal and vertical edge segments are assumed to be valid and subsequently retained. In the proposed model horizontal edges contain slopes between $\pm 2.5°$ and vertical edges contain slopes between $87.5°$ and $92.5°$. Positive and negative slopes are treated independently. Positive horizontal edge segments contain slopes greater than or equal to $0°$ and less than or equal to $2.5°$ and negative horizontal edge segments contain negative slopes greater than or equal to $-2.5°$ or equal to $0°$. The positive and negative dominant horizontal edges are determined by comparing and extracting the largest segments from the positive and negative horizontal edge segments, respectively. If either of the dominant horizontal edge segments have a slope of $0°$ only the larger of the two segments is considered as the dominant horizontal edge.

Positive vertical edge segments contain slopes greater than or equal to $87.5°$ and less than or equal to $90°$ and negative horizontal edge segments contain negative slopes less than or equal to $92.5°$ or equal to $90°$. The positive and negative dominant vertical edges are determined by comparing and extracting the largest segments from the positive and negative vertical edge segments, respectively. If either of the dominant vertical edge segments have a slope of $90°$ only the larger of the two segments is considered as the dominant vertical edge.

Images of natural outdoor scenes often contain a horizon. In these type of scenarios, the VP is often located on the horizon line. For an unsupervised algorithm it is not known a priori whether a scene is outdoor or indoor. However, irrespective of whether a horizon exists or not this line will nevertheless represent a dominant horizontal edge. As such, in the proposed model, an estimation of the horizon line is attempted for all scenarios. If both the dominant and non-dominant horizontal edges are interrogated, then the horizon line is considered to be the row containing the most energy. To account for slope inconsistencies a buffer of two rows above and below the respective row is considered i.e. the energy contained with five consecutive rows are analysed for each candidate row. The horizon is assumed to span a significant width of the image. From experiment any candidate row containing a nonzero pixel count of less than 30% of the width of the image is excluded from analysis. For the purpose of VP detection, the horizon edge is chosen as the largest connected component within the candidate horizon row. The horizon edge may but not necessarily be one of the positive or negative dominant horizontal edges.

Converging diagonal lines play the most significant role in the estimation of the VP. In the proposed model these types of lines are handled differently compared to horizontal and vertical straight lines. Owing to the ill-posed nature of the problem it is not possible to know a priori which

of the parallel diagonal lines provide optimal VP convergence. Moreover, to minimise over-clustering of the converging diagonal lines and subsequent skewing of the accumulator matrix it is necessary to optimise the collection of valid straight lines segments through the discarding of redundancies. Owing to the possibility of inherent discontinuities it is not feasible to remove all non-dominant redundant parallel diagonal lines. Therefore, in the proposed model, a more reserved approach is chosen.

For every segment there exists an associated perpendicular range. This may be visualised as the bidirectional protraction of every point on the straight-line segment in a perpendicular direction to the segment and across the entire range of the coordinate space. Initially two segments are considered to be parallel if their slopes are both positive or both negative and are within a close degree proximity of each other. Subsequently if the smaller segment intersects the perpendicular range of the larger segment, then the smaller segment is considered to be redundant and invalidated. This refinement is necessary for the minimisation of possible infinities as well as to account for the redundant intersections. From experiment a parallel degree range of 2° and a perpendicular overlap range of 80% is chosen. The dominant horizontal, vertical and horizon edges are always considered to be valid irrespective of the refinement process. An illustrative interpretation of the refined straight lines is shown in Fig. 68(c).

3) Candidate Vanishing Points

The candidate VPs are extrapolated from an associated accumulator array (AA). The AA may be described as a two-dimensional integer array. In the proposed model the AA canvas is expanded to account for candidate VPs that may potentially lie outside of the boundaries of the original image domain. From experiment the expansion is chosen to be two times the height and three times the width of the source image. If the ratio of the width to the height is less than 80%, then the expansion is chosen to be three times the height and five times the width of the source image. In either scenario the original spatial region is located vertically on the bottom and horizontally in the middle of the expanded domain.

The values in the AA are determined by extrapolating the intersecting points of every straight-line segment with every other straight line segment through solving of the polynomial from their respective slope-intercept forms. A value of 1 is added to the accumulator at each intersection point (row and column location) for each line pair. Illustrative interpretations of the intersecting refined straight lines are shown in Fig. 69(b).

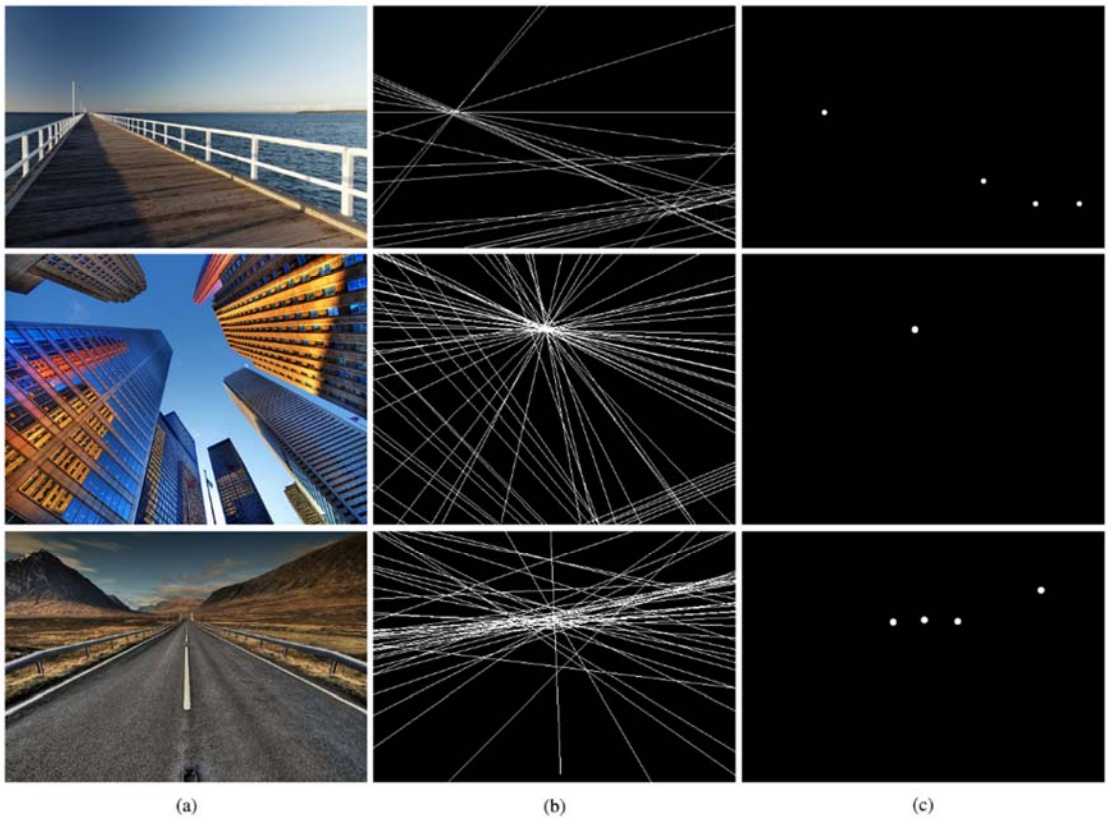


Fig. 69 Candidate VPs. (a) Original images; (b) Protracted straight lines (for illustrative purpose only); (c) Candidate VPs.

Each pixel in a digital image represents the quantisation of an associated point in 3D *analogue* space. As such each edge and subsequent straight line interpretation (especially the slope) of the edge will be prone to the errors associated with quantisation. Moreover, the extrapolation of the intersecting points of the quantised straight line segments are based on analogue polynomials. As a consequence, the maxima points in the AA may not necessarily represent the optimal candidate VPs. To address these concerns, the proposed model instead estimates the candidate VPs by evaluating the perpendicular distance between each non-zero location in the AA and every refined valid straight line segment. For each non-zero location in the AA the number of segments within a perpendicular range of p, $2p$ and $4p$ are determined. From experiment, p is chosen as 0.25% of the perimeter of the image.

In the proposed model up to 5 candidate VPs are extrapolated. Initially, from experiment, a candidate VP is chosen as the point in AA containing the maximum number of segments within a perpendicular range sorted in the order $2p$, p and $4p$. Subsequently, to avoid clustering of the candidate VPs, a $n \times n$ region centred on the respective candidate VP in the AA is set to zero prior to selecting the next candidate VP. Based on experiment, n is chosen as 2.5% of the perimeter of the image. This candidate VPs are illustrated in in Fig. 69(c).

4) Optimal Vanishing Point Estimation

Finally, from experiment, the dominant VP is chosen as the candidate VP containing the maximum number segments within a perpendicular range sorted in the order 4*p*, *p* and 2*p*. The segments with a perpendicular range of 4*p* are chosen as the associated vanishing lines (VLs). From experiment, *p* is chosen as 0.25% of the perimeter of the image. This is illustrated in Fig. 70.

Fig. 70 Dominant VP and associated VLs.

D. Results and Discussion

The results are divided into two sections. The first section discusses the training data used in the proposed model and the second section involves the testing of the algorithm against data not available during the training phase. Although the proposed model relies only on the straight edges in the image to estimate the VP, the algorithm is nevertheless multivariate. In order to establish a benchmark, supervised training of the algorithm is performed using the data associated with the previously proposed models. Subsequently, testing of the model is conducted. The testing component of the results section is vital to establishing the efficacy as well as the novelty of the proposed model.

1) Benchmarking

The training dataset consist of images from previous works as well as numerous random images from the WWW. These include the 6 images in Fig. 71 from Battiato et. al [23, 56], 3 images in Fig. 72 and Fig. 73 from Schaffalitzky and Zisserman [59], 3 images in Fig. 74 from Chappero et. al [57], 10 images from the York urban database provided by Denis et. al [58] and 33 images from the WWW (refer to Fig. 75). The York urban database is a collection of 102 images of indoor and outdoor urban structures and includes the associated ground-truths for the dominant straight line edges based on the Manhattan world assumption [164].

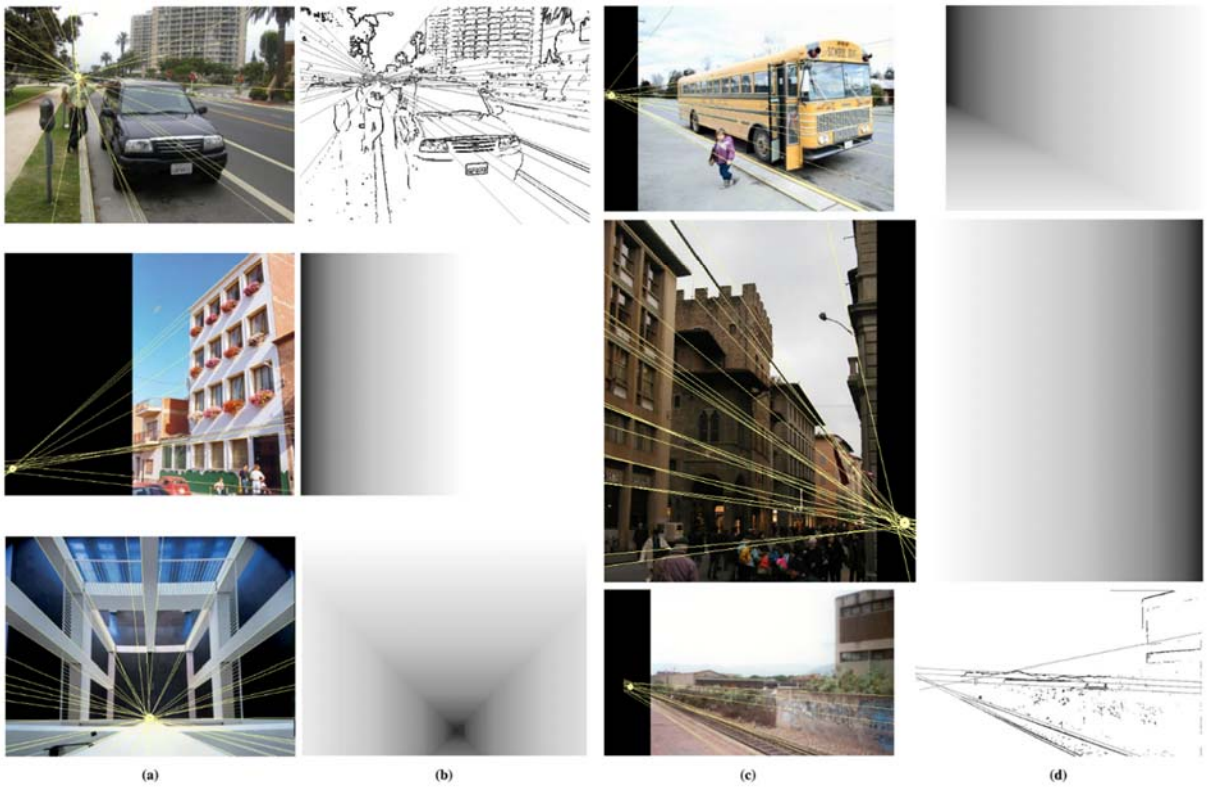


Fig. 71 Training set 1. (a) and (c) Results of proposed method; (b) and (d) Comparative results of the method proposed by Battiato et. al [23] using VLs and geometric depth maps.

A subjective comparison between the proposed method and the approach proposed by Battiato et. al [23] is provided in Fig. 71. Based on these results the proposed approach is comparable to the method proposed by Battiato et. al. Their method employs the use of the Hough transform accumulator technique for the detection of both the main straight lines as well as the intersections of the main straight lines. One disadvantage of the method proposed by Battiato et. al is that only concurrent lines are considered and horizontal, vertical and horizon lines as well as redundant parallel coplanar lines are not taken into account. Another disadvantage is that the HT accumulation matrix for straight line detection is constructed on a pixel-by-pixel basis compared to the more computationally efficient segment-based method employed in the proposed model. The method proposed by Battiato et. al also attempts to account for more natural environment type scenarios by incorporating a sort of rudimentary region-based approach using mean shift segmentation.

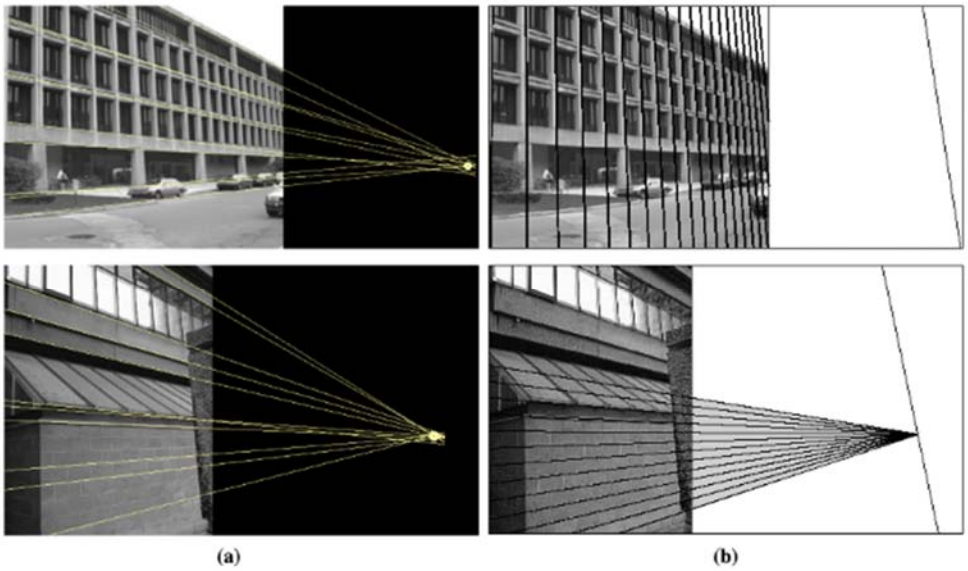


Fig. 72 Training set 2. (a) Results of proposed method. (b) Results from Schaffalitzky and Zisserman [59].

A subjective comparison between the proposed method and the approach proposed by Schaffalitzky and Zisserman [59] is provided in Fig. 72. Based on results the proposed approach is comparable to their method. Both methods extrapolate the straight-line segments directly from the edge map in the image plane parametric space and both consider the concurrent or converging lines to be a dominant factor in VP estimation. The significant difference between the proposed method and the Schaffalitzky and Zisserman approach is that the latter method considers coplanar equispaced parallel straight lines as another dominant influence in VP estimation. These heuristics, which includes features such as tiles, windows and planks on boardwalks or peers, may be useful in projecting possible VP regions, provided there exists a priori knowledge of the scene. However, in the case of unsupervised VP detection this information is unavailable.

Further inconsistencies may be highlighted through interrogation of the scenario evidenced in Fig. 73(b). In the Schaffalitzky and Zisserman method the brick wall may firstly result in conflicting horizontal and vertical coplanar parallel lines and secondly, create unnecessary clusters of straight line segment that may cause inconsistencies in the projection of the VP candidates. Another concern with the Schaffalitzky and Zisserman approach is that in addition to the detection of the parallel straight lines being performed in the parametric space using classical techniques their method also performs the detection using more elaborate feature extraction techniques.

In the proposed model consecutive (coplanar perpendicular space) parallel straight lines are assumed to be redundant and therefore only the dominant segment in each consecutive parallel straight line grouping is considered as a potential VL. The refinement of these line groupings is illustrated in Fig. 73(d) as well as Fig. 68(c). The added mathematical complexity in the Schaffalitzky and Zisserman approach appears to be unnecessary, since comparable, if not better,

results are achievable with the simpler proposed method. In addition to being computationally more efficient, the proposed approach requires noticeably fewer heuristics for the estimation of the VP.

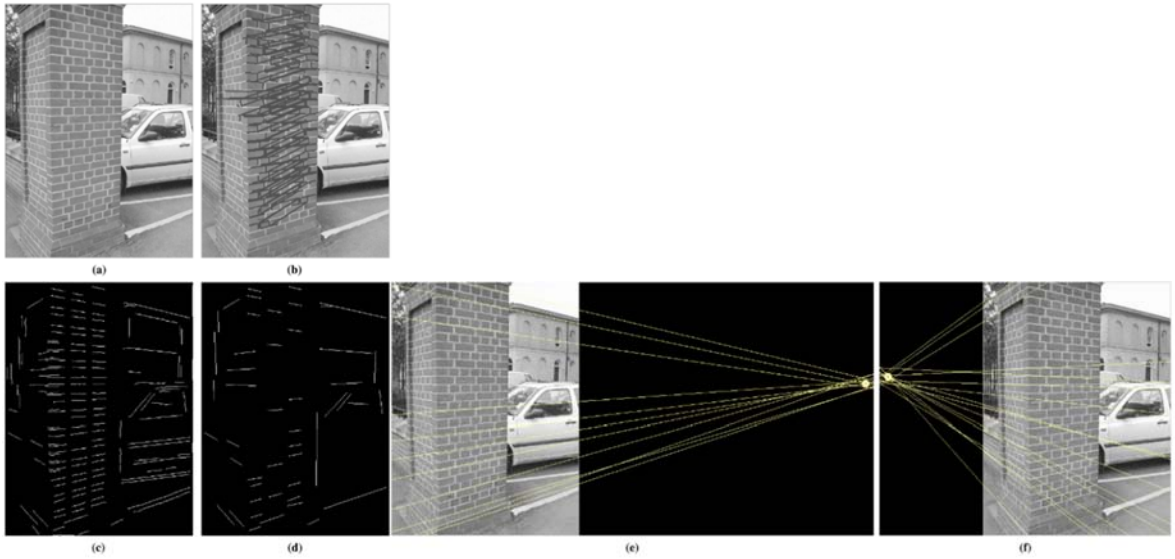


Fig. 73 Training set 2. (a) Original image; (b) Incorporation of coplanar parallel lines using the feature extraction method proposed by Schaffalitzky and Zisserman [59]; (c) Extracted straight edges using the proposed method; (d) Refined straight edges using the proposed method; (e) Results of the proposed method (Dominant VP); (f) Results of the proposed method (Secondary VP).

A subjective comparison between the proposed method and the approach proposed by Chappero et. al [57] is provided in Fig. 74. Based on these results the proposed approach is comparable to method proposed by Chappero et. al. However, the significant advantage of the proposed method is that it is dependent on only one criteria (edge segments) as opposed to the three (Hough, pixel gradients and segmentation) considered by Chappero et. al. In addition, the latter method requires two accumulation matrices to be constructed and analysed; one for the Hough transform and the other for the first derivatives (gradients) in the image plane. Another disadvantage is that the accumulation matrices are constructed on a pixel-by-pixel basis compared to the more computationally efficient segment-based method employed in the proposed model. In the method proposed by Chappero et. al an attempt is also made to account for more natural environment type scenarios by incorporating a sort of rudimentary region-based approach using mean shift segmentation; this is similar in a way to the aforementioned Battiato et. al approach. The proposed model also takes into account VPs that may lie outside of the parametric space; this eventuality is discussed but not implemented in the approach proposed by Chappero et. al.

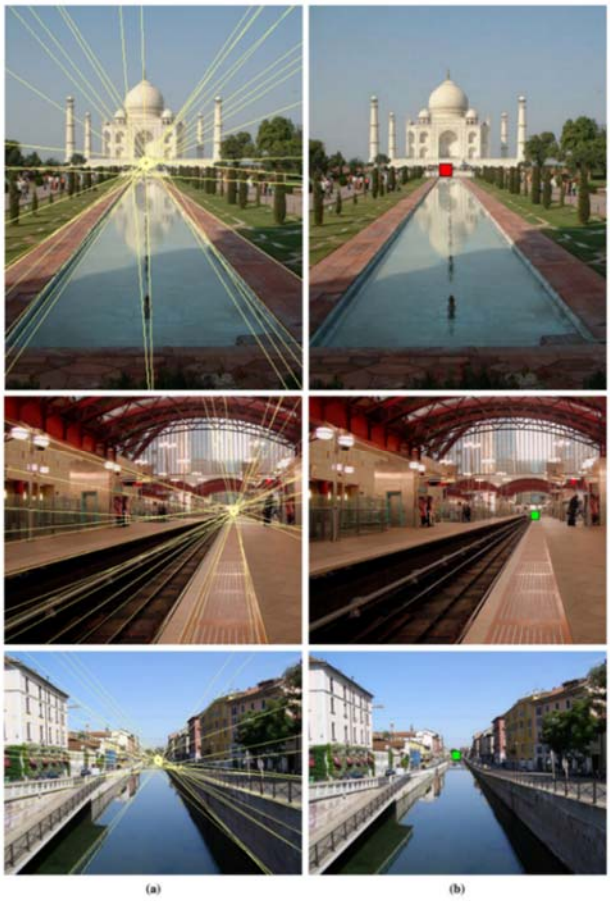


Fig. 74 Training set 4. (a) Results of proposed method; (b) Results from Chappero et. al [57].

Owing to space limitations only a limited number of images from the training dataset are provided in this section. The complete database of training images and associated results are available from the author at serenr@gmail.com.

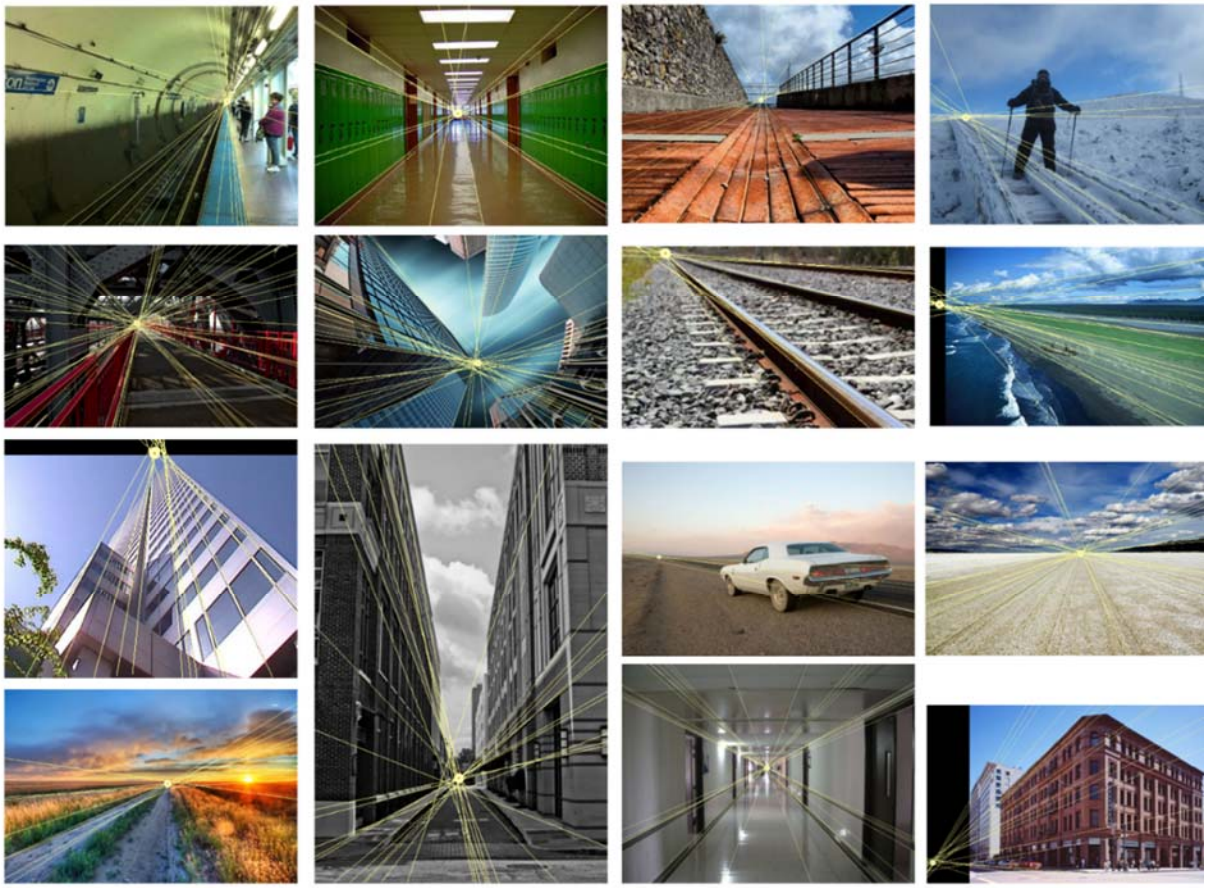


Fig. 75 Training set 1. 15 of the 33 training images from the WWW.

2) Testing

The testing dataset consists of 144 images representing a variety of VP scenarios. As with the training images the test images chosen provide a reasonable amount of variation such as indoor and outdoor man-made structures as well as natural environments. In addition, attributes such as colour, illumination, texture, size etc. are also taken into account. These include 52 images from the WWW (refer to Fig. 76) and 93 images from the York urban database (refer to Fig. 77). The latter also includes the associated ground-truth VP coordinates.

Owing to space limitations only a limited number of comparative results of the total dataset of test images are presented in this section. The complete database of test and training images and the associated results is available from the author at serenr@gmail.com.

In the proposed model the VP is considered to be the dominant convergence point used to guide the observer's 3D depth perspective. Although this point is imaginary, it nevertheless

represents a point in 3D space at a location in the distant of the scene in the direction of the imaginary perspective lines.

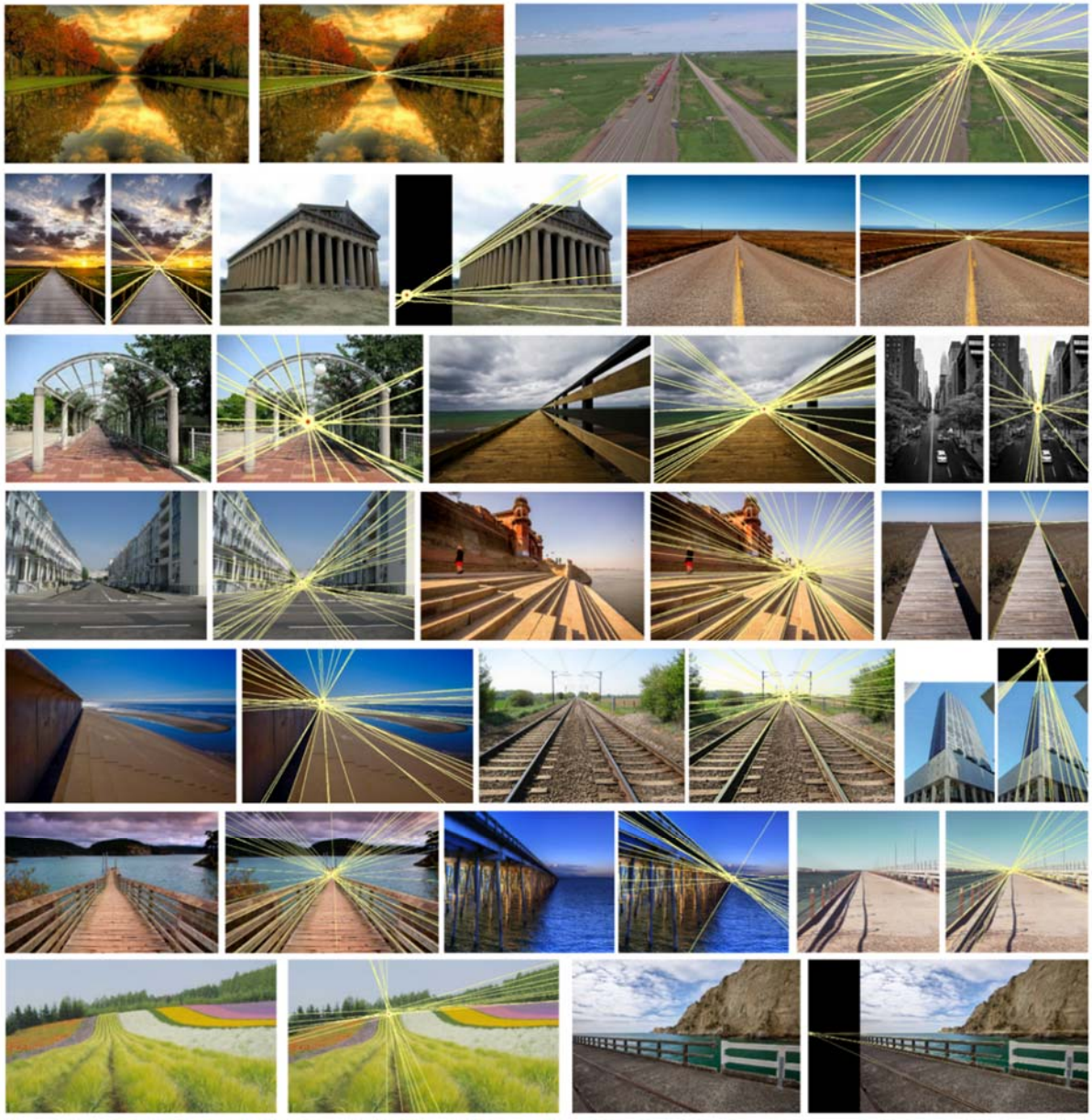


Fig. 76 Test results using the proposed method. The images pairs are numbered 1 to 19 starting with 1 on the top left corner and moving column-wise then row-wise until 19 on the bottom right corner.

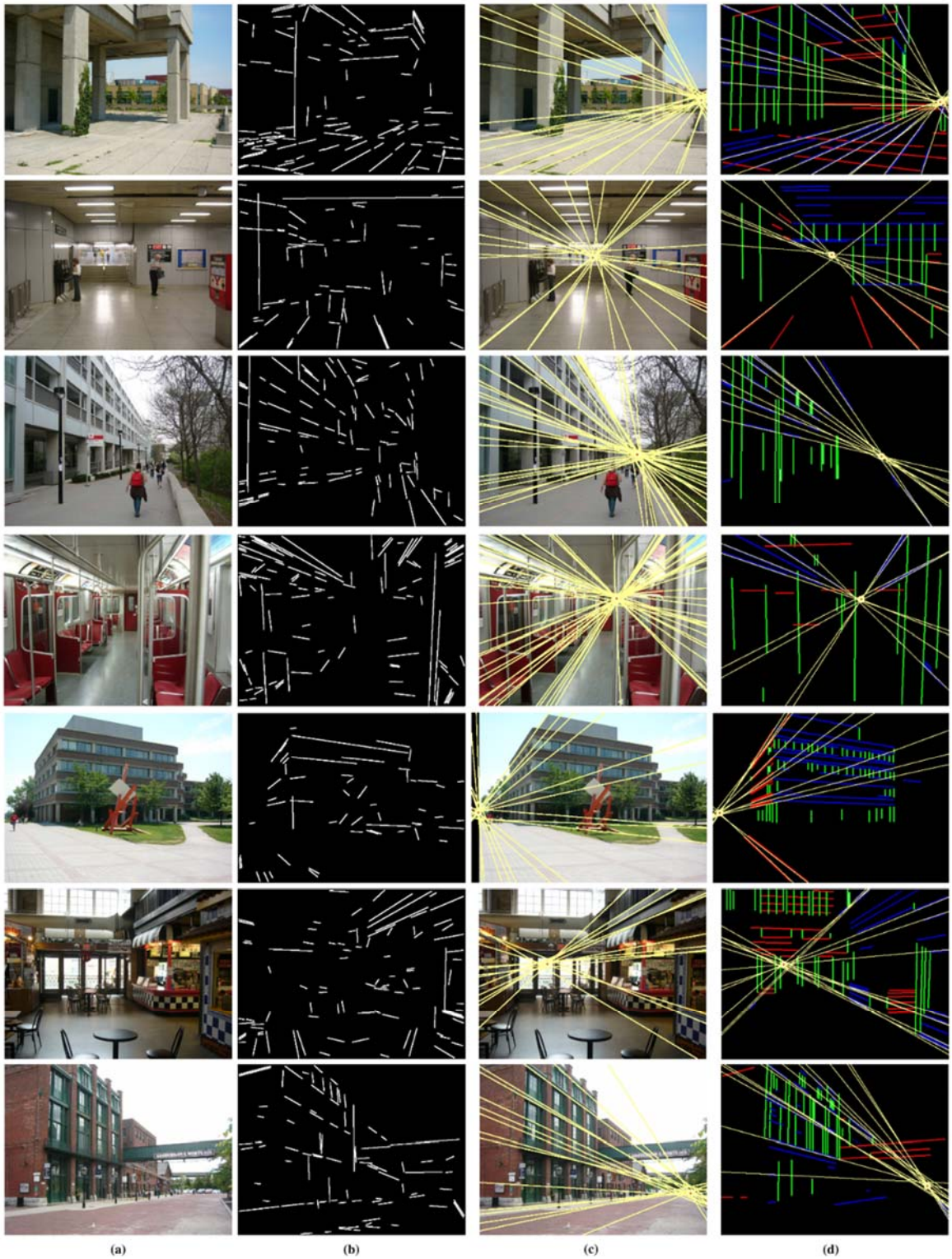


Fig. 77 York urban database. (a) Original image; (b) Refined straight edges extracted using the proposed method; (c) Results of proposed method; (d) Results of proposed method applied to the respective York urban ground-truth.

The following is a brief analysis of the results shown in Fig. 76. Image 15 (second row from bottom, first column from the left) shows that according to the proposed model the VP is estimated to be the convergence region at the end of the peer. However, one may subjectively consider the VP to be located on the horizon line at the base of the mountains in the background. The reason for the VP being located at the extrapolated point is that man-made structures like the peer exhibit more dominant digital straightness compared to natural structures such as mountains and trees. Another example is given in image 14 (third row from bottom, first column from the right); in this case the estimated VP is considered to be in the direction of the apex of the building where the VLs converge to a point outside the parametric space. While this appears to be subjectively correct, in this scenario there are actually two more VPs (east and west) in addition to the projected VP (north). The proposed model is also tested against more natural environments; such as image 18 (last row, first column from the left). In this case the striations between the rows of plants together with the dominant horizon line present on the ridge of the field result in VP being located on the horizon line but nevertheless in the centre of the dominant convergence region.

The proposed model is also evaluated using the source and ground-truth data provided by the York urban database. This dataset comprises both outdoor and indoor images of man-made environments. Some of the results are illustrated in Fig. 77.

A success rate of 81% (75 out of 93) is achieved when compared to the ground-truth data. The accuracy of the data is based on a buffer of 1% of the sum of the width and height of the source image around the ground-truth VP.

The limitation of the proposed method for VP estimation is that it is inherently dependent on extrapolation of the straight lines in the image. This is a concern for images of natural environments where minimal straight lines are present. Further research is warranted in this regard.

E. Conclusion

A novel unsupervised method is proposed for estimation of the VP in images containing artificial or man-made environments. The simplicity of the proposed approach is that it is only dependent on the analysis of a few edge segments, which are extracted from a binarised edge map, directly in the image-plane coordinate space. Moreover, the proposed method achieves similar, if not better, accuracy than other methods that involve the use of more mathematically complex techniques such as Gaussian spheres, polar-space transformations, Helmholtz probability, Bayesian inference, Hough transforms and segmentation.

The proposed model consists of four stages. Firstly, the straight edges in the image are estimated. The results show that horizontal, vertical and diagonal straight edges may be extrapolated directly from the edge map of the image by evaluating the energy present within the

confines of each connected component. Secondly, redundant straight line edges are excluded. The results show that two types of redundant edges may be excluded. The first are the non-dominant overlapping parallel edges and the second are non-dominant parallel edges within the perpendicular co-planar space of the dominant edges. The third stage of the proposed model involves the selection of candidate VPs. The results show that these candidate VPs may be extrapolated firstly, determining the intersecting points of every straight-line segment with every other straight line segment through solving of the polynomial from their respective slope-intercept forms and secondly, using a block-based approach to evaluate the straight lines passing within a perpendicular buffer distance of each intersecting point. The fourth and final stage involves the extraction of the optimal VP from the selection of candidate VPs. The results show that the optimal VP may be determined by performing a more refined interrogation of vanishing lines that are within a perpendicular distance of the candidate VPs.

Future work may include the expansion of the model to account for natural environments using a type of region-based approach. The vanishing regions that converge towards a VP tend to be triangular in nature. By exploiting these properties, it may be possible to infer straight lines and subsequently extrapolate a VP. However, for this type of approach it may be necessary to incorporate a method to discern between artificial and natural environments.

VI. STEREOSCOPIC IMAGE SYNTHESIS OF SINGLE 2D LOW DEPTH-OF-FIELD IMAGES USING DEPTH IMAGE-BASED RENDERING

A. Introduction

The purpose of a 2D-to-3D conversion system is to output two images representing a source image or video frame from two different viewpoints. Depth image-based rendering (DIBR) is the term given to the process of generating or rather synthesising images of a scene that appear to be simultaneously captured from different viewpoints of that scene using only a single source image and its associated depth map [39]. In the case of 2D-to-3D conversion only two of these simulated viewpoints are required. However, if the original source image is chosen as one of the viewpoints, then only one synthesised disparity image is actually needed.

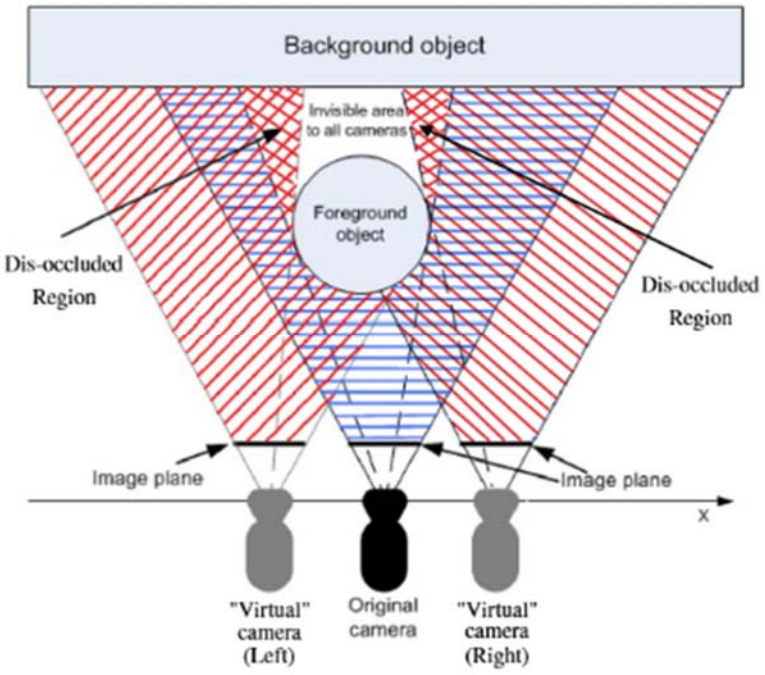


Fig. 78 Extrapolation scenario of virtual cameras views from original camera (Image adapted from [165])

To create these so-called "virtual" views, pixels representing objects closer to the camera, as determined by the depth map, need to be moved either left or right, in front of the background pixels. This pixel location reassignment process is commonly referred to as *warping*. A

consequence of warping is some pixels will become occluded and other pixels will be revealed, resulting in missing pixels or "gaps" appearing in the background, as shown in Fig. 78.

The problem with these dis-occluded regions is even with a highly accurate depth map there is no precise means of determining the values of the revealed pixels, since the depth map only provides depth information and not intensity nor lateral nor angular information of the occluded regions. In most cases, specifically single images and static video frames, these values may only ever be predicted within a certain probability, since they never actually exist. These newly created holes need to be resolved effectively so as not to significantly distort the quality of the generated disparity images. This sub-step of the DIBR process, which involves the predicting and subsequent filling, of these missing pixels is referred to as hole-filling.

It is common to set the original image as one view, say the right eye and then synthesize the left-view. The advantage is it would only be necessary to generate one additional view. However, the disadvantage will be the size of the dis-occlusion spaces occurring in the rendered view. By synthesising two views these missing pixel spaces may effectively be halved since it will be split across both images. The fundamental problem with these missing pixel spaces is that there is no information in the original reference image or depth map that may be used to accurately determine the values of the dis-occluded regions. In essence, the information needed may only ever be obtained from an image that had been simultaneously captured from the new viewpoint, for example, by using a stereoscopic camera. This is not an easy problem to solve and owing to these irregularities directly affecting the subjective image quality, the hole-filling problem is the most significant issue currently existing in DIBR.

This paper is organised as follows: Section B provides a brief discussion of some of the related work. Section C provides a description of the proposed model as well as the necessary steps required to autonomously generate two disparity views of a single 2D LDOF image using DIBR. Section D reports the experimental results. Section E provides a brief summary of the research and closes the paper.

B. Related Work

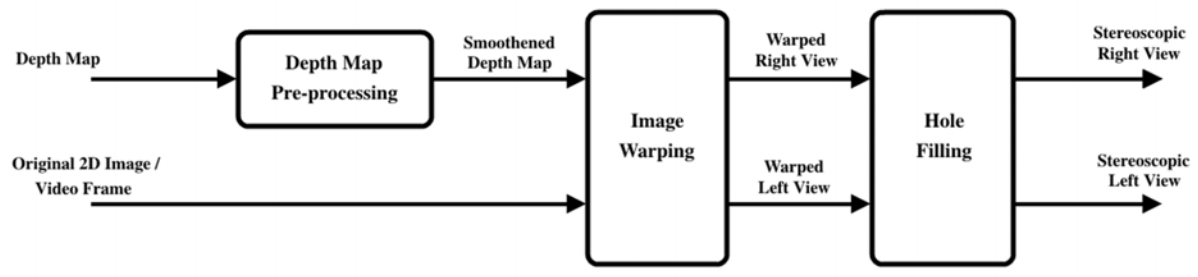


Fig. 79 Depth image-based rendering (DIBR) System (Image adapted from [47])

There are three primary processes involved in DIBR. These include depth map pre-processing, 3D image warping and hole-filling [47]. A typical DIBR system is illustrated in Fig. 79. *Depth map pre-processing* involves the possible prior smoothing of the depth map as well as the setting of certain camera and geometric constraints. The *image warping* component is responsible for creating the disparity shift between the left and right images. The *hole-filling* element addresses the assignment of pixel intensity values to the missing (dis-occluded) pixel spaces occurring as a consequence of the image warping.

The following section provides a deeper discussion into some of the techniques previously proposed for the different stages of the DIBR process. 3D image warping is discussed prior to depth map pre-processing in order to provide a deeper understanding for the latter requirement.

1) 3D Image Warping

By using the original 2D source image or video frame as a reference and *relatively* shifting the pixels either left or right, according to the depth information provided in the depth map, different images of the reference scene may be created in such a way so as to appear as though they have been simultaneously captured by cameras located either slightly to the left or right of the original camera position. This pixel shifting process employed to generate these alternative viewpoints is commonly referred to as 3D warping or just *warping* [166]. The essence of 2D-to-3D conversion is to simulate stereoscopy and, therefore, only *two* viewpoints or "virtual" camera views are needed. In the case where the original image is chosen as one of the viewpoints, then for the "second eye" it is only necessary to synthesise *one* other viewpoint.

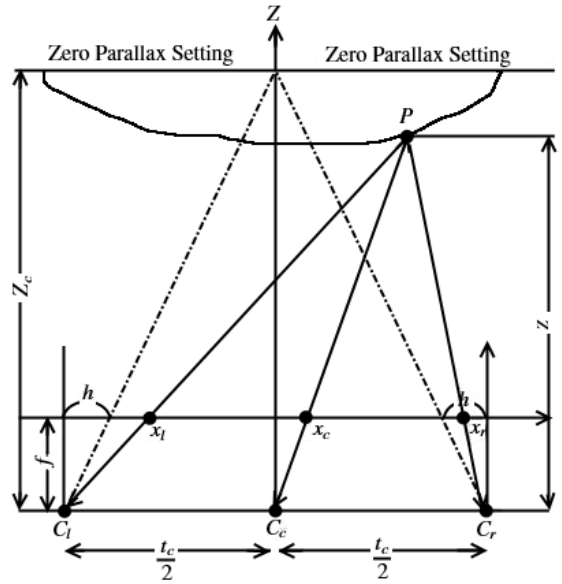


Fig. 80 Camera configuration for rendering of virtual stereoscopic images [48].

Image warping contains two steps. The first involves the projection of each pixel in the real view image into the 3D world based on the camera parameters and the second involves the re-projection of the pixels back to the 2D image of the virtual view for view generation [167]. Fig. 80 provides the geometrical basis of the DIBR warping process [47]. The left and right eye images, located at the "virtual" camera positions C_l and C_r, respectively, may be generated for a certain lateral distance t_c based on the focal length f and the depth map value Z. The geometrical relationships are expressed as

$$x_l = x_c + \frac{t_c \times f}{2} \cdot \left(\frac{1}{Z} - \frac{1}{Z_c} \right) \tag{40}$$

and

$$x_r = x_c - \frac{t_c \times f}{2} \cdot \left(\frac{1}{Z} - \frac{1}{Z_c} \right) \tag{41}$$

where Z_c, referred to as the *convergence distance*, is the distance between the camera and the *zero-parallax setting* (*ZPS*).

The product of t_c and f generates a value indicating the *maximum* disparity or lateral shift in the synthesised images. Although both t_c and f are constant for each scenario they are nevertheless unique to each scenario. Once each of the variables have been assigned a value, these warping equations become even more simplified and may, in a straightforward process, be used to directly map pixels for the left and right eye view, synthesising the stereoscopic image pair [47].

2) Depth Map Pre-Processing

Depth map pre-processing is the first step in the DIBR process, implying it is actually performed prior to the 3D warping stage. This stage usually involves the selection of the abovementioned convergence distance Z_c or *ZPS* and maximum disparity shift value, given by the product of t_c and f, as well as the smoothing of the depth map. Subsequently, warping may be performed by inputting Z_c, t_c, f and the modified depth map values Z directly into Eq. (40) and Eq. (41).

The relationship between disparity, depth and horizontal (lateral) position is illustrated in Fig. 81. Here it shown that only objects within the Z_{near} to Z_{far} range are perceived as 3D while those outside these limits are viewed as 2D. Incidentally, this explains why applications destined for 3DTV do not exceed the five-meter limit, since depth beyond this distance from the camera is not visually perceptible on these systems. Moreover, Fig. 81 shows that the scene will not change if the viewer decides to shift his point of view horizontally [168].

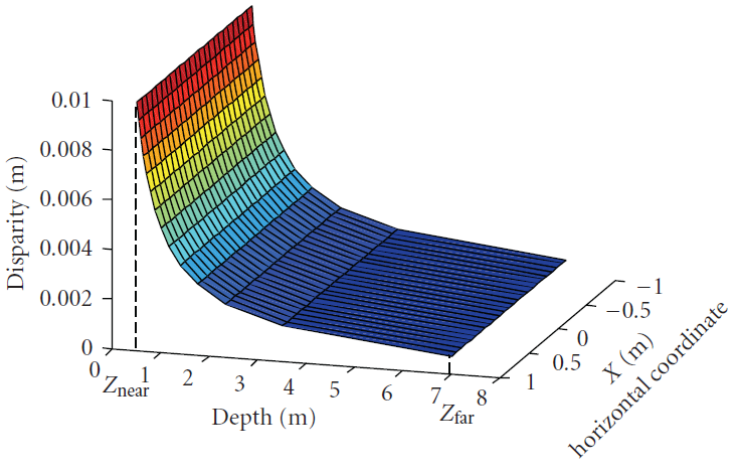


Fig. 81 Relationship between disparity, depth and horizontal position for sample parallel cameras ($t_c = 0.1$ m and $f = 0.05$ m) [168].

Although there exist several methods for estimating the *ZPS* [169], the "shifted" depth map approach is commonly preferred for 3DTV [47]. If Z_{near} denotes the depth value closest to the observer and Z_{far} denotes the depth value furthest from the observer, then

$$Z_c = \left(\frac{Z_{near} - Z_{far}}{2}\right). \tag{42}$$

For a typical 8-bit depth map the range will most likely be $Z_{near} = 0$ and $Z_{far} = 255$. The depth map will then be adjusted and *normalised* such that the Z values lie in the interval of $[-0.5, 0.5]$.

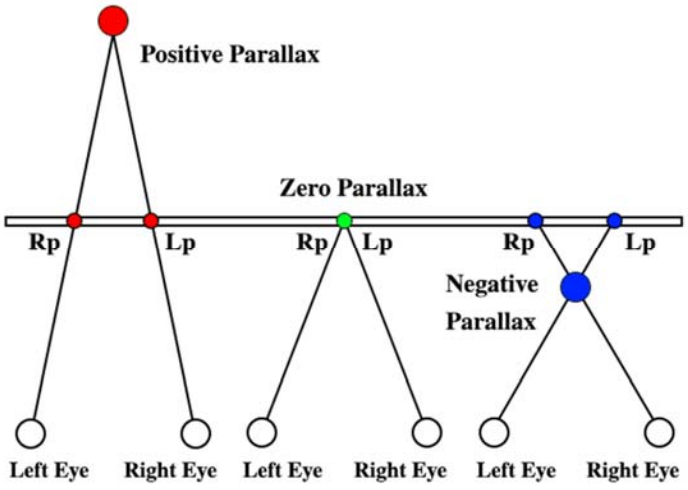


Fig. 82 Binocular parallax [170]

Three types of "binocular" parallax may be experienced by an observer. These are termed zero, positive and negative parallax [10]. Fig. 82 illustrates what may occur, for example, in 3DTV and cinema. Zero parallax has the eyes converged at the plane of the screen, positive parallax will result in the objects appearing submerged into the display screen and in the case of negative parallax, objects will appear as being between the observer and the display screen. Pixels located between $0 < Z \leq 0.5$ will result in positive parallax, while pixels located between $-0.5 \leq Z < 0$ will result in negative parallax.

For sporadic images it is unlikely the baseline distances t_c as well as camera parameters, such as *f*, are known a priori. Therefore, in the case of an autonomous 2D-to-3D conversion system, it will be required to formulate values for t_c and *f*. For a comfortable view a 1° horizontal disparity offset is suggested as the maximum depth range. For a standard 4×3 image the maximum lateral shift will translate to approximately 5% of the width of the image. In addition, a viewing distance of four times the image or video frame height is also recommended [47, 171].

The purpose of smoothing the depth map, which is usually performed through Gaussian filtering, is multi-fold. Firstly, it may be used to effectively reduce noise and blocky artefacts and secondly, redistribute inaccuracies and thirdly, smoothen transitions at object boundaries, across the entire depth map [172]. Moreover, it may be applied more aggressively to the edges of the OOI in order to alleviate or even remove some of the dis-occlusion distortions arising as a consequence of warping [39]. This smoothing function is discussed in more detail below, in the hole-filling subsection.

3) Hole-Filling

Fig. 83 DIBR Image Warping [47]. (a) Original image; (b) Warped "Left-view" image (The red indicates the dis-occluded regions).

A consequence of warping is that some pixels will become occluded and other pixels will be revealed, resulting in missing pixels or "gaps" appearing in the background. This is illustrated in Fig. 83. These newly created holes need to be resolved effectively so as not to significantly distort the quality of the generated disparity images. This sub-step of the DIBR process, which involves the prediction of the intensity values and subsequent filling of these missing pixels, is referred to as *hole-filling*.

As mentioned earlier, it is not uncommon to set the original image as one of the viewpoints, for example, the right eye and then synthesise only the left-view. The advantage is it that only one additional view is needed. However, the disadvantage will be the increased size of the resulting dis-occluded regions in the synthesised view. On the other hand, by synthesising two views these missing pixel spaces may effectively be halved, since it will be split across both images. Irrespective of whether a single view or two views are synthesised, the fundamental problem is that there is no information in the original reference image that may be used to precisely determine the intensity values of these missing pixels.

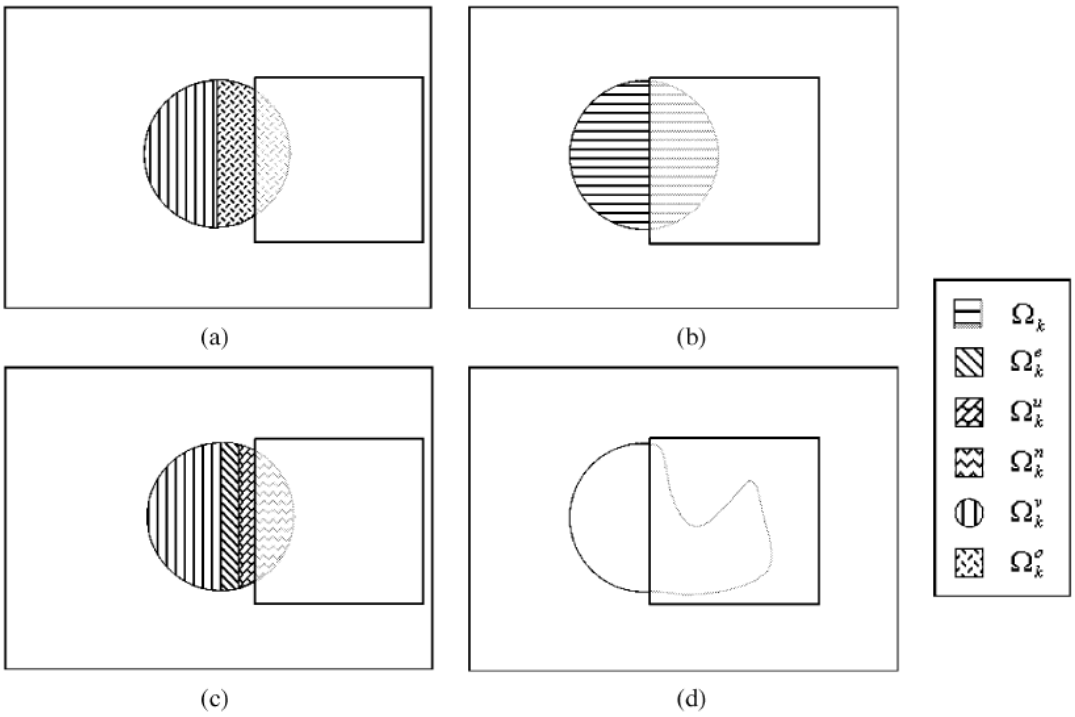


Fig. 84 Occlusion analysis [173]. (a) Left frame of a stereo pair (Synthesised); (b) Right frame of a stereo pair (Original); (c) Categorisation of occluded pixels; (d) Possible noncircular object occluded by the square (right frame).

The complication surrounding the missing pixel information may be more clearly understood by considering the scenario illustrated in Fig. 84. Assuming Fig. 84(b) represents the original image (right eye) and Fig. 84(a) represents the virtually generated or synthesised view (left eye),

then from this scenario, Fig. 84(b) cannot provide any texture information regarding the newly exposed pixels. Even with a highly accurate depth map there is still no precise means of determining the values of these revealed pixels, since a depth map only provides depth information and not lateral nor angular information of the dis-occluded regions. Moreover, as shown in Fig. 84(d), there is actually no way of telling whether the information behind the square is in fact circular in shape i.e. the information needed is only ever obtainable from an image simultaneously captured from the new viewpoint, for example, by using a stereoscopic camera.

Since the inception of DIBR, several approaches have been proposed to address the hole-filling problem [47-49, 166, 174, 175]. These techniques are grouped under two categories. The first is pixel interpolation- or extrapolation-based methods and the second is depth map smoothing methods

Interpolation or *extrapolation* techniques are commonly referred to as *image inpainting* [176]. Inpainting provides a means of filling these dis-occlusions by either mapping pixels from the original image or extrapolating or mirroring values, using the foreground (FG) or background (BG) information, onto the synthesised views [166]. The six most popular inpainting schemes employed in DIBR include [49, 174, 175]:

1) Constant colour filling. This involves estimation of the average colour of the closed boundary of each hole (FG and BG objects are not separated) and subsequent filling of the entire hole with a unique and constant colour.

2) Horizontal interpolation. This is a basic interpolation of the known boundary pixels in the horizontal direction. Each dis-occluded region is divided into single pixel height rows and each individual row is then filled by interpolating the intensity values of the endpoints left and right towards the centre. The relative depths of objects are not accounted for so the FG and BG objects will appear to be fused together.

3) Horizontal extrapolation with depth information. This method is similar in concept to the horizontal interpolation technique. However, in this case the relative depth of objects are taken into account. Since the 3D warping will only create dis-occluded regions in the BG, the boundary pixels of the foreground objects may essentially be ignored i.e. extrapolation of intensity information is based only on the border pixels constituting the objects in contact with the hole at the greatest relative depths. As with the standard horizontal interpolation approach, each dis-occluded region is split into single pixel rows. However, in this case only one endpoint is considered, resulting in interpolation either occurring from left to right or vice versa i.e. there is no joining or fusing of FG and BG objects.

4) Edge oriented averaging with depth information. This method is identical to the previously discussed method in regard to handling of the FG and BG boundary pixels. However, instead of performing direct horizontal interpolation of the background border pixels, interpolation is based

on the minimal intensity difference in four directions around each pixel. This technique attempts to improve and preserve the boundary transitions.

5) Diffussion-based interpolation. This is a more sophisticated method of interpolation than the previous techniques and is based on the smoothing and subsequent inward diffusion of the pixels on and near the hole boundaries. This technique may nevertheless be considered as an extension of the first two approaches. There are many variations to diffusion-based interpolation such as linear, nonlinear (for example, Laplacian), isotropic and anisotropic. A specific model may be chosen depending on the properties of the local neighbourhood such as direction or structure curvature. In general diffusion-based interpolation approaches have shown to exhibit less perceptual inconsistencies such as abrupt disconnected edges and border transitions.

6) Exemplar-based interpolation. These approaches may be seen as an amalgamation of the aforementioned methods together with texture synthesis techniques [177, 178]. Inspiration is taken from local region-growing techniques whereby attempts are made to grow the texture of the local neighbourhood one pixel or one patch (referred to as *exemplar*) at a time. The dis-occluded regions may be synthesised by learning and sampling from similar regions (in the form of random textures or regular patterns) in known parts of an image or a texture sample. Exemplar-based interpolation approaches are more suitable to supervised models.

In general, exemplar-based followed by diffusion-based image inpainting techniques are considered to be the most popular. The former is more applicable to the recovery of texture of large areas while the latter is more suited to small regions as well as straight lines and curves. Although these two interpolation techniques are significantly more sophisticated than the other aforementioned approaches they are nevertheless more accurate and robust.

Although there have been major advances in the field of image inpainting these methods may nevertheless suffer from discrepancies such as stripe distortions, abrupt boundary transitions as well as other texture artefacts. As a means of ameliorating or even removing these inconsistencies, depth map smoothing has been proposed [47, 166]. These smoothing techniques, which are briefly discussed in the previous subsection, are not actually dis-occlusion filling or inpainting methods but rather are considered as a means of either reducing or altogether avoiding the existence of dis-occlusion regions through the smoothing of the depth map prior to 3D warping.

Although there are several filter types that may be applied to achieve smoothing of the depth map, for conceptual simplicity the Gaussian filter $h(n, \sigma_n)$ is considered. This is defined as

$$h(n, \sigma_n) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{n^2}{\sigma_n^2}\right), \tag{43}$$

for $-\frac{w}{2} \leq n \leq \frac{w}{2}$, where n implies the direction (either horizontal, μ, or vertical, υ), σ denotes the depth smoothing strength and w represents the filter's window size. Given a depth map d

where $d(x, y)$ indicates the depth of the pixel in the associated original image at position (x, y), then the depth value after smoothing $\hat{d}(x, y)$ will be equal to

$$\hat{d}(x,y) = \frac{\sum_{\upsilon=-\frac{w}{2}}^{\frac{w}{2}} \left\{ \sum_{\mu=-\frac{w}{2}}^{\frac{w}{2}} \left(d(x-\mu, y-\upsilon) h(\mu, \sigma_\mu) \right) h(\upsilon, \sigma_\upsilon) \right\}}{\sum_{\upsilon=-\frac{w}{2}}^{\frac{w}{2}} \left\{ \sum_{\mu=-\frac{w}{2}}^{\frac{w}{2}} h(\mu, \sigma_\mu) h(\upsilon, \sigma_\upsilon) \right\}}. \tag{44}$$

Manipulation of w and σ will have a direct influence on the 3D warping process, whereby the dis-occluded regions will either be reduced in size or eliminated altogether, depending on the degree of smoothing. In both cases, this will have a relatively positive impact on the subjective quality of the synthesised images. There are two types of smoothing that may be performed; these include symmetric and asymmetric. For the former, $\sigma_\mu = \sigma_\upsilon$, and for the latter, $\sigma_\mu \neq \sigma_\upsilon$. Fig. 85 provides an illustration of both forms of depth map smoothing.

Although symmetrical smoothing has shown to minimise or even eliminate the aforementioned texture artefacts associated with interpolation, the synthesised images nevertheless are prone to geometric distortions, in particular when object boundaries are vertically straight. However, noticeable improvements to these distortions are observed when smoothing is performed *asymmetrically*, specifically in the cases where $\sigma_\mu \ll \sigma_\upsilon$ [47, 167]. Asymmetric smoothing, with emphasis in the vertical direction, is characteristically congruent with the binocular modus operandi of the HVS, whereby horizontal shifts rather than vertical disparities are responsible for depth perception [179]. This type of asymmetric filtering provides smoothing of the sharp depth changes, in particular around the border transitions of the OOIs, while still providing reasonably good and stable disparity depth cues.

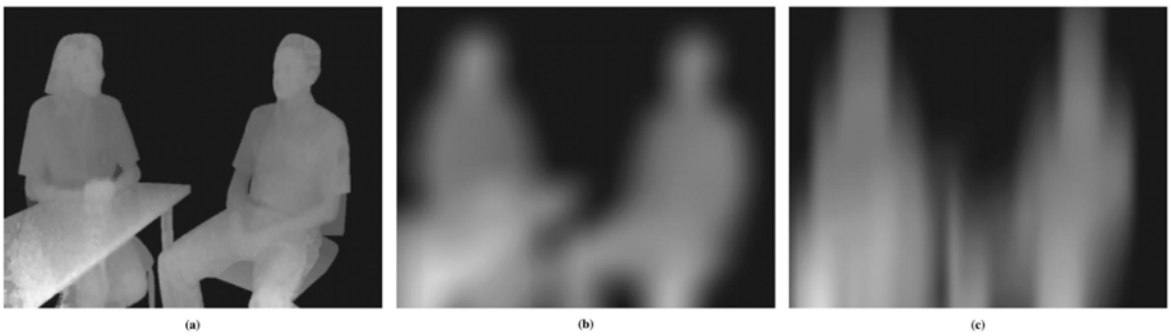


Fig. 85 Depth Map Smoothing [47]. (a) Original Depth Map (The higher the intensity value implies the pixel is located closer to the camera i.e. a shallower depth); (b) Symmetrically smoothed depth map; (c) Asymmetrically smoothed depth map (with a stronger emphasis in the vertical direction).

Another type of filter, often considered, which is an adaptation of the Gaussian filter, is the bilateral edge-preserving filter [128, 129]. This is described on *p*. 95 together with a discussion on the adaptations and improvements to these types of filters.

All methods that attempt to address the inconsistencies associated with hole-filling contain some inherent discrepancies and most scenarios are prone to non-ideal conditions. As a consequence, in this research a combination of these techniques, in conjunction with other approaches, are considered as means of achieving reliable and autonomous DIBR.

C. Proposed Approach

The purpose of this stage is to produce two images representing two slightly different lateral viewpoints of the original source image i.e. a left and right disparity image-pair. In the proposed model this is achieved through the application of depth image-based rendering (DIBR). As discussed above, DIBR is divided into three stages. These include depth map pre-processing, image warping and hole-filling.

1) Depth Map Pre-processing

The purpose of the depth map pre-processing substage is to assign the convergence distance and maximum disparity shift value as well as smoothen the depth map.

For viewer comfort, the convergence distance is set to $[-0.5, 0.5]$. This is denoted by Z_c in Eq. (40) and (41) on *p*. 137. The effect of setting $Z_c = 0.5$ is that the eyes will converge on the screen at the 127-128 depth range i.e. values below 127 (shallower depth) will appear to pop out of the screen towards the viewer and values above 128 (deeper depth) will appear to be submerged into the screen away from the viewer.

In the proposed model the original image is not considered as one of the disparity pair. By rendering both the left and right disparity images the dis-occlusion distances created during the 3D warping process are effectively reduced by half. The maximum disparity shift value is set to 10% the width of the image; this is denoted as $t_c \times f$ in Eq. (40) and (41) on *p*. 137. Since the dis-occlusion distances are distributed across the left and right rendered disparity pair there will be a maximum horizontal shift of 5% of the width of the source image for each disparity image.

To further alleviate some of the dis-occlusions generated during the image warping substage, an asymmetrical bilateral filter (BF), instead of a Gaussian, is applied to the depth map. The choice of edge preserving may appear to be counterintuitive since the purpose of the smoothing is to spread the effect of the warping in a manner so as to minimise the resulting dis-occlusions. However, the problem with not preserving the edges, especially of the OOI, is that border regions of the objects are also spread out during warping resulting in perceptual irregularities appearing

in the synthesised images. The choice of the BF therefore allows for the spreading of the warping to be performed primarily in the regions between objects.

Fig. 86 Gradient-based depth maps. (a) Original images; (b) Depth maps; (c) Asymmetrically smoothed depth maps; (d) Final merged asymmetrically smoothed depth maps.

In this research the joint shiftable filter proposed by Chaudhury et. al [129, 134, 135] is chosen; this is discussed on *p*. 95. From experiment, a spatial standard deviation $\sigma_s = 0.55\%$ the perimeter of the image, a range standard deviation $\sigma_r = 1.7\sigma_s$ and a $n \times m$ smoothing window with $n = \sigma_s$ and $m = 5n$ is chosen for the BF. This is illustrated in Fig. 86(c).

The final pre-processed depth map is produced by merging the filtered depth map with the original depth map and using the shallower depth value as the dominant intensity. To ensure that the integrity of the OOI remains intact after smoothing, particularly around the border regions, the entire OOI region is allocated the same depth value. This is illustrated in Fig. 86(d).

2) 3D Image Warping

In this research both the left and right disparity images are rendered by applying the method proposed by Zhang and Tam [47]. This is described by Eq. (40) and (41) on p. 137. In the proposed model $t_c \times f = 10\%$ the width of the image (compared to 5%), $Z_c = 0.5$, x_c denotes the current column value of pixel $x(r, c)$ and Z indicates the normalised value of pre-processed depth map at location $x(r, c)$; x_l and x_r will indicated the shifted column value of $x(r, c)$ in the disparity image pair.

Fig. 87 3D warping. (a) Original image; (b) Disparity warp using original depth map; (c) Disparity warp using the asymmetrically smoothed depth map.

The shifting of pixels during the image warping process will result in several location conflicts. In these cases the pixel with the smallest Z value (shallower depth) will take precedence. This implies that the objects and regions with shallower depth values will partly or fully occlude (overlap and replace) the objects and regions with deeper depth values at the same location. Moreover, depending on the location and depth of objects and regions, the shifting process may result in several dis-occluded (blank) regions being generated in the disparity images. Furthermore, pixels on the left and right border that are shifted to the right and left, respectively, will also result in blank regions. This is illustrated in Fig. 87(b). As discussed above, by pre-processing the depth map with an asymmetrical filter prior to warping, results in the reduction of the spacing between the gaps and in some instances the complete resolution of the gaps. This is illustrated in Fig. 87(c).

3) Hole-Filling

The consequence of the warping process results in the appearance of dis-occlusions. For the disparity images to be effective, these inconsistencies need to be adequately resolved. In other words, the holes need to be filled. In the proposed model hole-filling or inpainting is achieved by firstly, performing depth map smoothing prior to 3D warping (discussed earlier on *p.* 144) and secondly, closing the dis-occlusions through horizontal interpolation using depth as well as the direction of warp and thirdly, applying localised smoothing to some of the filled regions. The combination of directional edge-based horizontal interpolation and diffusion-based inpainting is chosen over the exemplar-based approaches owing firstly, to the small sizes of the disocclusions and secondly, to the blurring effect being more closely associated with the low-frequency textural information usually attributed to the background regions in LDOF images and thirdly, to the former being more appropriate for unsupervised applications. The techniques applied for hole-filling are the same for both the left and right disparity images, except for the reversal in direction.

Dis-occlusions occur when objects or regions at a shallower depth are moved left or right in front of object or regions at a deeper depth. By this understanding, the dis-occlusions represent discontinuities in the objects or regions at the deeper depth. As a consequence, the colour intensity chosen for the horizontal interpolation is based on the occluded object or region.

The setting of the convergence distance to $[-0.5, 0.5]$ results in objects and regions being shifted in both the left and right directions within each of warped disparity images. Moreover, the effect of the depth map smoothing creates interspersed dis-occlusions within a larger dis-occlusion region. As a consequence, the direction of warp is required in determining which of the several possible occluded regions is the source of the horizontal interpolation. If a dis-occlusion is created by an object or region being shifted to the left, then the horizontal interpolation of the dis-occlusion should be from left to right and vice versa for an object or region being shifted to the right. Any remaining gaps are filled using vertical interpolation. In this research a method referred to as interpolation averaging is also proposed. In this case a second horizontal-vertical interpolated image is produced using a new set of intensity values as the reference. From experiment the pixels located at a quantised distance of two away from the border pixels are chosen. Subsequently, the final interpolated image is chosen as the average of the two interpolated images.

The interpolation process may result in the appearance of blocky lateral striations. This is illustrated in Fig. 88(b). To minimise these stripy artefacts, additional smoothing is performed. Only large regions adjacent to the defocussed BG regions are considered for additional smoothing since these regions are firstly, associated with the larger dis-occlusions and secondly, are expected to blend in with the smoothed appearance of the defocussed BG regions. The defocussed regions in an image may be described as being Gaussian in nature. Therefore, in the proposed the smoothing is performed primarily using Gaussian filtering.

From experiment, a region is considered to be large if it either contains at least 50 non-zero elements and has a horizontal distance greater than m pixels or has as vertical distance greater than m pixels, where $m = 0.8\%$ of the width of the image.

Fig. 88 Stereoscopic images. (a) Warped disparity image; (b) Hole-filling using horizontal interpolation.

Smoothing is performed in two stages. For the first stage smoothing is applied to two regions of the large gaps. These include the inner and outer regions. The inner region is determined by eroding the large gaps. From experiment, a 5×5 square structuring element is chosen for the erosion. This will result in the removal of the outer two boundary layers from the large gap regions. For these inner regions, the final interpolated image together with a buffer around the region are initially smoothed using a Gaussian low-pass filter. Subsequently pixels associated with the inner region are delineated. From experiment a buffer of 2 pixels is chosen and for the filter a 9×9 square matrix and a standard deviation of 1 is chosen.

The outer region is determined by dilating the large gaps and subsequently removing the inner region. From experiment, a 3×3 square structuring element is chosen for the dilation. This will result in a region that may be described as a hollow close bounded segment with a boundary thickness of 3 pixels. For these outer regions, the final interpolated image together with a buffer around the region are initially smoothed using a circular average filter. Subsequently pixels

associated with the outer region are delineated. From experiment a buffer of 2 pixels is chosen and for the filter a radius of 5 is chosen.

Fig. 89 Stereoscopic images. (a) Original image; (b) Final left disparity image; (c) Final right disparity image.

From experiment, the image is shown to exhibit less inconsistencies when the intensity of the perimeter pixels of both the smoothed inner and outer regions are set as the original value from the final interpolated image. The aim of the smoothing is to be able to match the defocussed BG region to some extent. Based on a significant portion of the test images, adjusting the filter parameters according to the size of the image is not necessarily more effective.

Minor irregularities may occur on the border transition owing to the smoothing performed on inner and outer regions of the large gaps. As a consequence, a second smoothing stage is considered in the proposed model. This smoothing is performed on a thin region between the border of the OOI and the newly interpolated and smoothed regions, from the first stage, using an averaging filter. Initially smoothing is applied to the entire image. However, only the intensities associated with a single pixel thickness region located at a pixel distance of 2 from the perimeter of the OOI are replaced in the final rendered image. The final synthesised left and right disparity images are illustrated in Fig. 89.

This concludes the proposed hole-filling sub-process of the DIBR process and subsequently concludes the model proposed for the autonomous rendering of a stereoscopic image pair from a single 2D LDOF image.

D. Results and Discussion

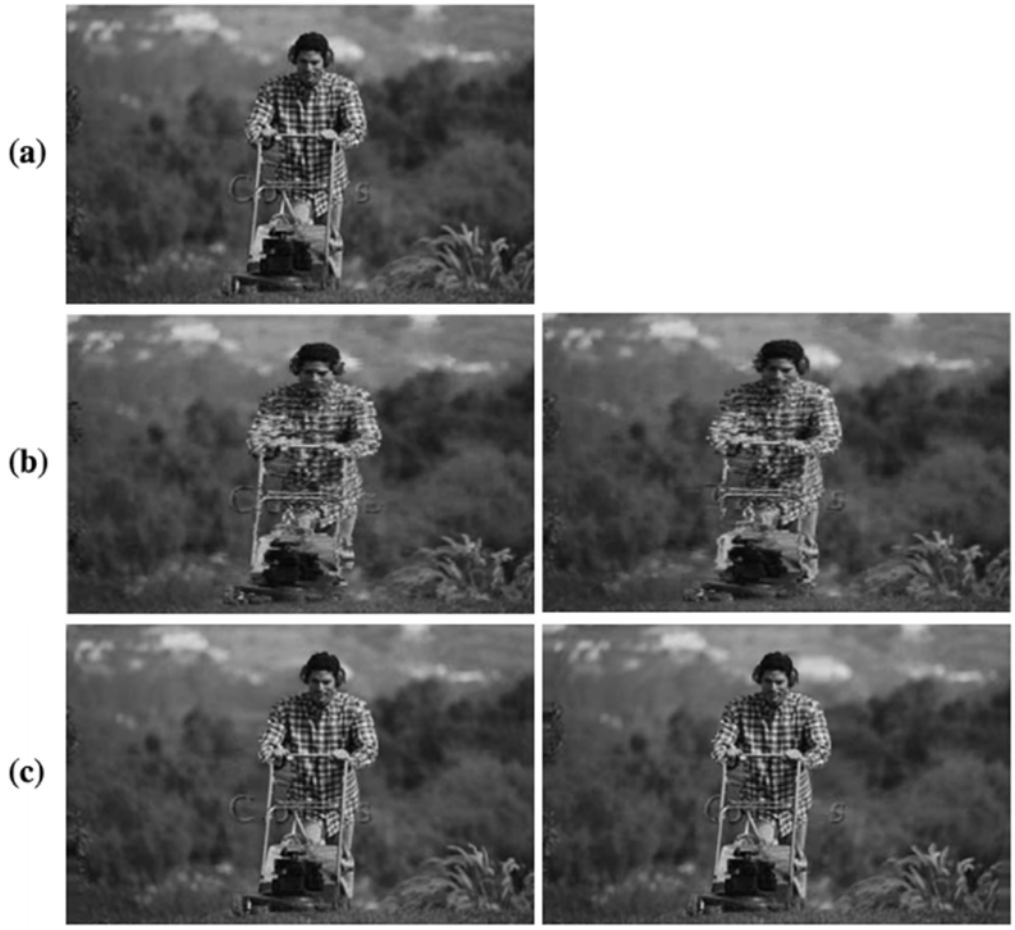


Fig. 90 Subjective comparison 1. (a) Original image; (b) Left and right views generated by Valencia et. al [62]; (c) Left and right views generated using the proposed method.

The training and test data are sourced from previous works as well as numerous images from the World Wide Web (WWW). Comparative analyses are performed against two other proposed methods [62, 65].

A set of comparative test results is illustrated in Fig. 90. The original image is illustrated in Fig. 90(a). Fig. 90(b) shows the results obtained from the method proposed by Valencia et. al [62]. Fig. 90(c) present the result of the proposed method.

When considering the method proposed by Valencia et. al, the hole-filling is achieved by the interpolation or stretching of the image. The concern with this approach is that the OOI in the synthesised images represent an augmented version of the original object. Although this distortion resolves the dis-occlusions, the skewing of the regions in the synthesised images makes this approach highly inaccurate.

Even though there exist some minor inconsistencies in the proposed method, it nevertheless deals more accurately with the dis-occlusions. Moreover, the entire original closed boundary OOI remains intact in both the synthesised images.

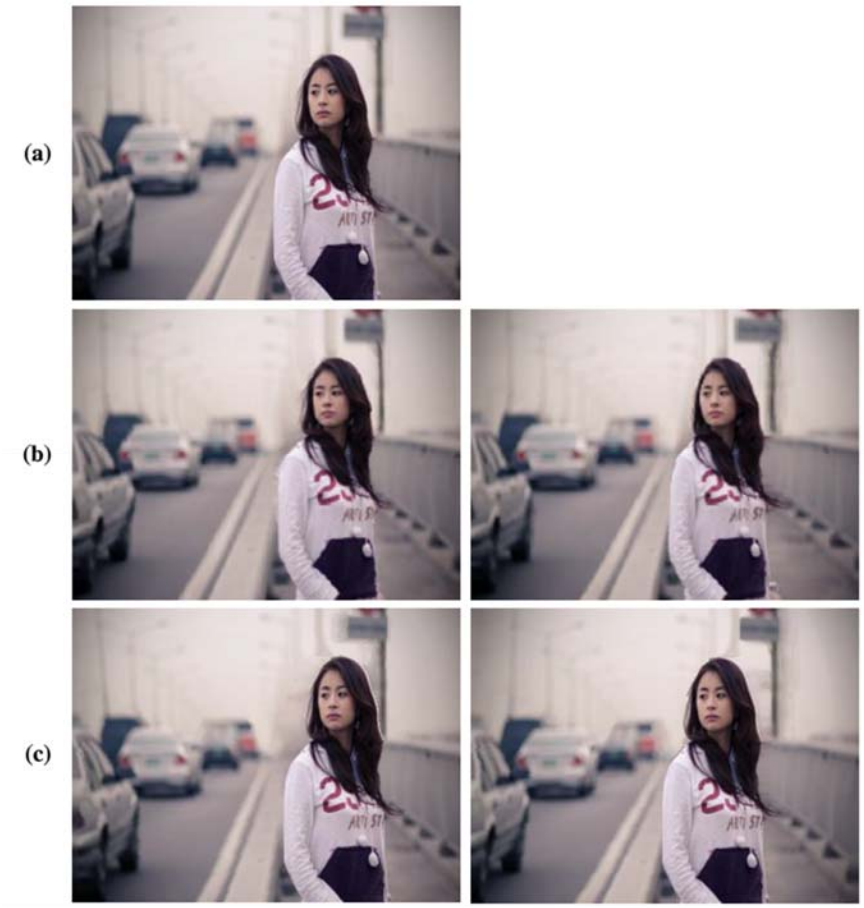


Fig. 91 Subjective comparison 2. (a) Original image; (b) Left and right views generated by Ko et. al [65]; (c) Left and right views generated using the proposed method.

A second set of comparative test results is illustrated in Fig. 91. The original image is illustrated in Fig. 91(a). Fig. 91(b) show the result obtained from the method proposed by Ko et. al [65]. Fig. 91(c) present the result of the proposed method.

When considering the method proposed by Ko et. al, although the entire closed boundary OOI remains intact, there are parts of the background that are inconsistent in terms of 3D projection. For the left image, the object (girl) should be shifted to the right. Although this is done with regards to the closed boundary region of the OOI, the concern is with the background regions. For example, the sign post is located some distance further away from the girl and therefore the shift of the girl left or right is expected to either widen or narrow the gap, respectively, between the girl and the sign post. However, this is not the case. Another problem occurs in the image of the left view. In this instance the regions near the right arm of the girl appear to be inconsistent. This is owing to asymmetric smoothing of the depth map where the constant depth value of the entire OOI is not retained.

Although there exist some minor inconsistencies in the proposed method, it firstly is comparable to the method proposed by Ko et. al and secondly, deals more accurately with the background regions in the synthesised images.

Additional results for some of the synthesised disparity image pairs for the training and test images are provided in Fig. 92 – Fig. 94.

There are some apparent limitations to the proposed approach. Firstly, the efficacy of the 3D warping of the objects and regions in the synthesised images is inherently dependent on the accuracy of the depth map in terms of both the delineation as well as the relative depth values. The delineation is important since the objects and regions are skewed left and right in the synthesised views based on the relative depth values associated with the individual objects and regions in the original image.

When considering the OOI, for example, if the delineation falls within the boundaries of the OOI, then parts of the OOI will be misaligned, misplaced or missing in the synthesised images. For example, the butterfly's antennae in the synthesised *Butterfly* images in Fig. 93. Another concern associated with this scenario is that there may be additional discontinuities or inconsistencies associated with the interpolation. For the hole-filling to be effective only the background pixels should be interpolated. However, in this scenario since the parts of the OOI outside the delineated region will be considered as belonging to the background. As a consequence, these in focus OOI pixels might be incorrectly interpolated together with the non-OOI pixels, which are often associated with defocussed regions. Although the smoothing of the depth map may alleviate some of these inconsistencies, in some cases the effect is nevertheless noticeable; for example in the region near the right ear of stag in *Stag* image in Fig. 94.

Conversely, if the delineation exceeds the boundary of the OOI then parts of the background will be misaligned, misplaced or missing; for example, the background in-between the wisps of hair on the right-hand side of the man's head in left synthesised *Forever* image in Fig. 94.

Secondly, the efficacy of the smoothing of the interpolated gaps is dependent on the degree of defocus of the adjacent regions. When there is a high degree of defocus in the background the smoothing of the interpolated regions appear to be consistent with the adjacent background. Examples of these type of scenarios is the upper defocussed regions in *Kid* image in Fig. 92 and the non-OOI regions in the *Girl* and *Butterfly* images in Fig. 93 as well as the *Forever* image in Fig. 94.

If the adjacent pixels exhibit a high degree of focus, particularly in the case of type 2 and 3 classifications, then the smoothing actually degrades the regions surrounding the OOI. Examples of these include the regions around the legs of *Kid* image in Fig. 92 as well as the non-OOI region in the *Girl* and *Butterfly* images in Fig. 93. These inconsistencies may also apply to images with relatively medium degrees of defocus, such as the *Bird* image in Fig. 92.

A possible resolution to address the perceptual inconsistencies associated with filling and smoothing of the dis-occluded regions adjacent to the high- and medium frequency textured background regions may be to incorporate unsupervised exemplar-based interpolation together with some sort of adaptive smoothing that varies the filter window size according to depth. Further research is warranted in this regard.

Fig. 92 Synthesised stereoscopic image pairs of selected training images: *Kid* (top) and *Bird* (bottom)

Fig. 93 Synthesised stereoscopic image pairs of selected test images – Set 1: *Girl* (top) and *Butterfly* (bottom)

Fig. 94 Synthesised stereoscopic image pairs of selected test images – Set 2: *Stag* (top) and *Forever* (bottom)

Owing to space constraints only a selected number of the stereoscopic image results are presented in this section. The appendix on *p*. 188 contains a selection of uncompressed or in some instances, optimally resized, training and test results. The author may be contacted at serenr@gmail.com for more than 300 LDOF images and their associated synthesised disparity image pairs.

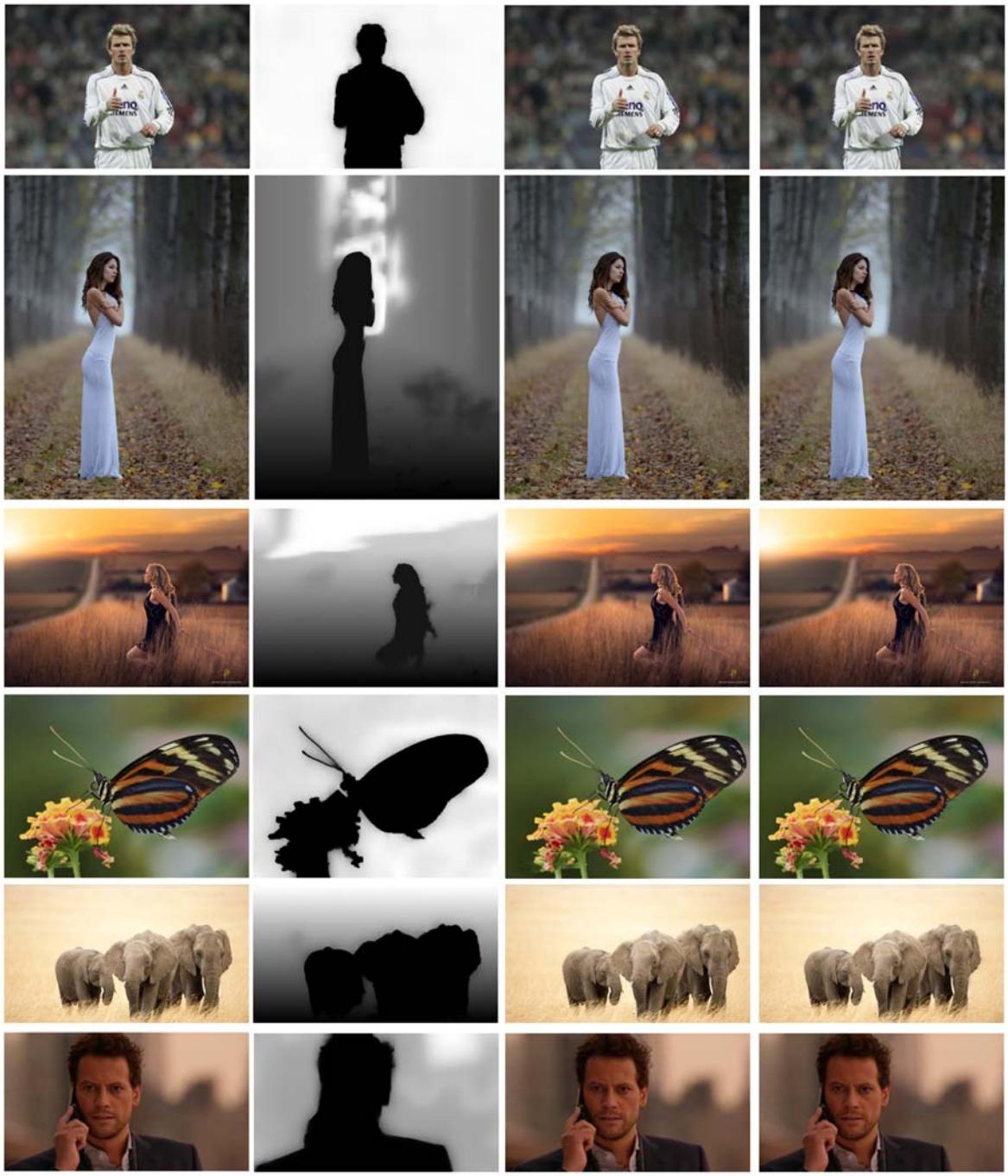


Fig. 95 Results of synthesised stereoscopic images – Set 1.

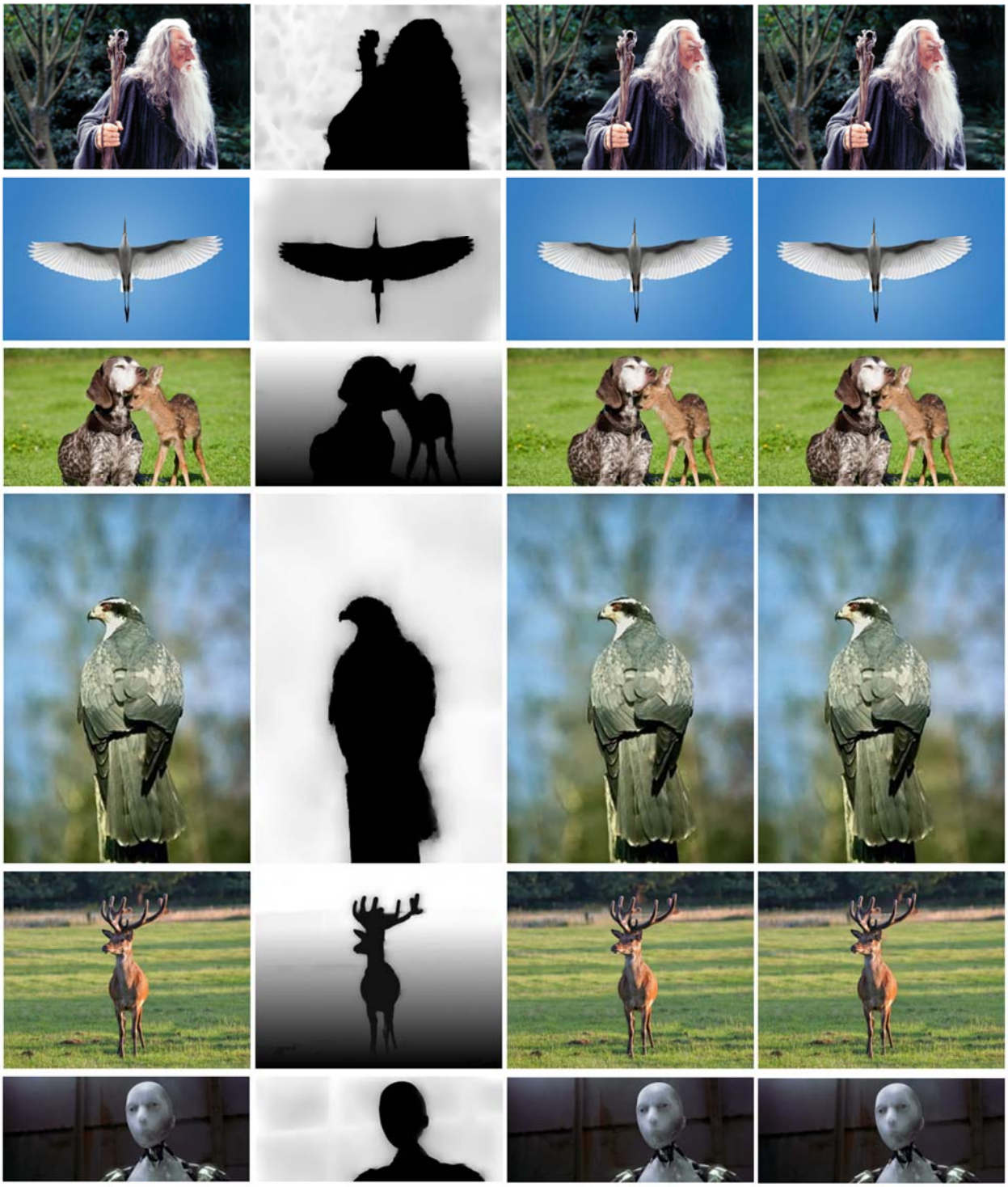


Fig. 96 Results of synthesised stereoscopic images – Set 2.

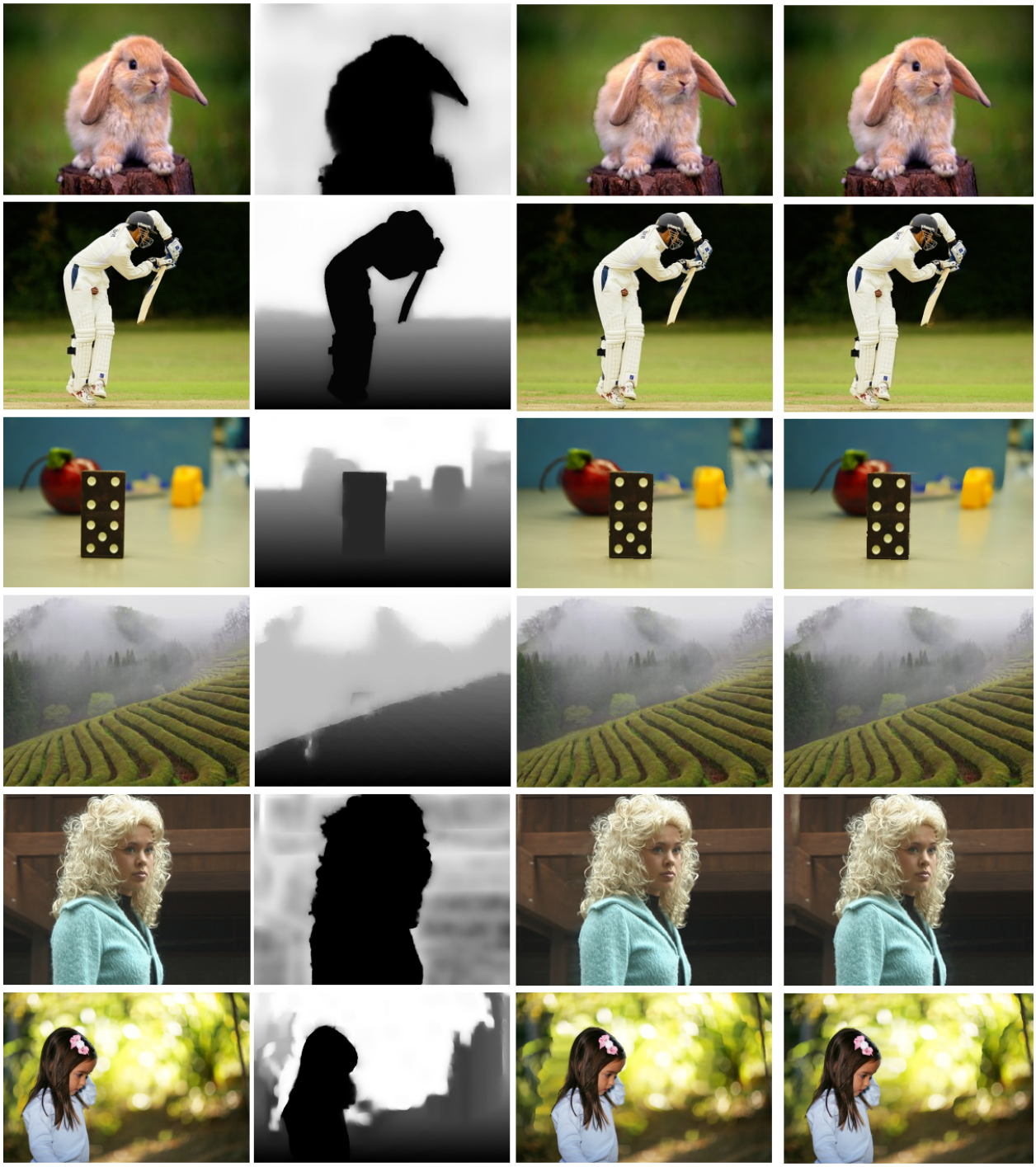


Fig. 97 Results of synthesised stereoscopic images – Set 3.

E. Conclusion

A novel unsupervised method is proposed for the synthesis of two images that represent a scene being viewed from two different perspectives. The stereoscopic images are produced by firstly, warping or laterally shifting the objects and regions according to a depth map of the LDOF image and secondly, resolving the dis-occlusion through the use of interpolation and low pass filtering.

The results show that by applying asymmetrical smoothing prior to warping reduces the degree of dis-occlusions. Moreover, the remaining inconsistencies may be effectively resolved by performing combined horizontal-vertical interpolation of the background regions that takes direction of warp into account together with smoothing of the new interpolated regions. The results show that the blocky striations within the interpolated regions may be effectively minimised or eradicated by combining Gaussian and circular averaging filtering together with average filtering on the border transitions. This makes it possible to automatically generate extremely accurate stereoscopic images of single 2D LDOF images.

Future work may include more precise hole-filling techniques by incorporating automatic exemplar-based approaches as well as using machine learning and Bayesian inference and the expansion of the proposed method to account for high DOF images.

VII. 3D (ANAGLYPH) IMAGE GENERATION

A. *Introduction*

There currently exist four popular presentation modes employed for the display of 3D images and video. These include anaglyph, polarised, shutter and autostereoscopic; the first three are illustrated in Fig. 98 [180].

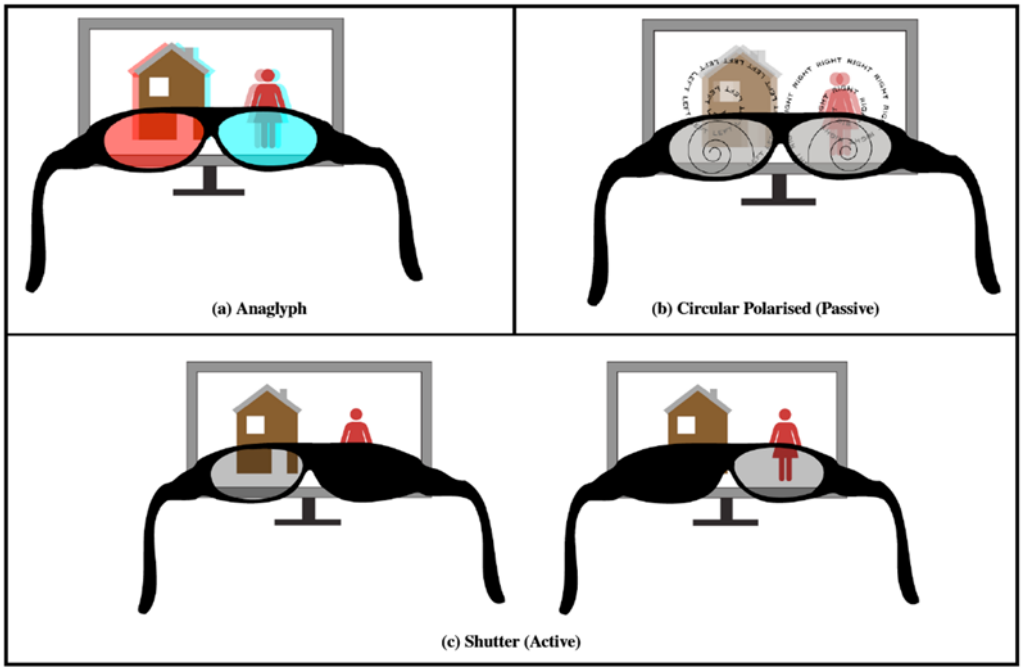


Fig. 98 3D Display Formats [180]

Anaglyph images are produced by superimposing two individually coloured (typically red and cyan) disparity images into a single image. This image may then be viewed through glasses having similar coloured filters.

In the case of polarised, each of the disparity images are individually circularly polarised and displayed simultaneously. Usually the left eye is polarised clockwise and the right eye is polarised anticlockwise and glasses containing circular polarising filters for each eye may then be used to view these images. These glasses are also referred to as "passive" because no power source is required for their operation. They are relatively cheap to manufacture and the most popular mode for 3D cinema. Circular polarisation is not a digital effect, whereby the disparity images are fused and stored as a single tangible digital image, as is in the case of anaglyphs, but rather the physical effect of the filters of the projection system simultaneously projecting the disparity images onto the screen [181].

For the shutter format, active glasses are used to alternatively "black-out" each eye depending whether the respective left or right image is being displayed on the screen. The glasses contain LCD lenses that "shutter" in complete synchronisation with the respective disparity images. The shutter effect is essentially unnoticeable owing to the extremely rapid speed at which

the shuttering occurs. These glasses are referred to as “active”, owing to the power source needed to operate the synchronisation sensor and LCD lenses. They are relatively expensive and largely reserved for use with 3DTV. The choice of “active” vs. “passive” has to do with the resolution of the images, since in the former the 3D image is essentially half the resolution of the 2D image quality being displayed.

Autostereoscopic displays refer to any 3D display set up not requiring any special glasses; in this technology the display itself acts as the disparity filter. These displays usually also incorporate some type of eye tracking technology that allows for the individual left and right disparity images to be simultaneously targeted to each retina of the observer.

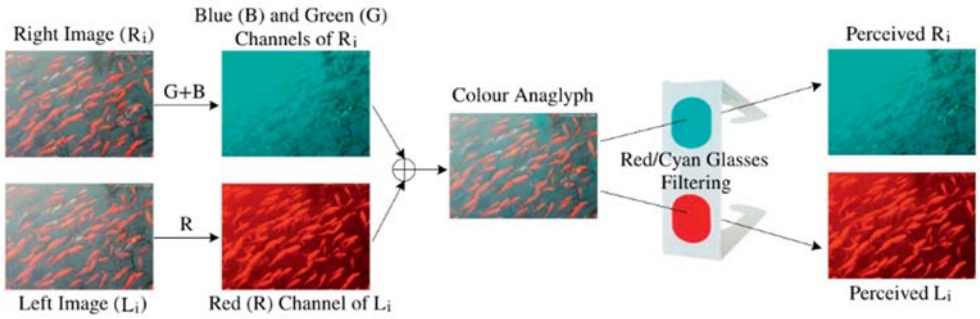


Fig. 99 Red-cyan anaglyph 3D displaying [50]

Of these four approaches, anaglyph is the simplest and most economical [50]. The anaglyph image concept was introduced in 1853 by Rollman [182], with the first printed anaglyph being attributed to Ducos du Hauron in 1891 [183]. This allows for the perception of depth through the splitting of the red, green and blue (RGB) colour channels across a stereoscopic image pair. The separation of the colour channels and subsequent superimposition of the stereoscopic pair into a single anaglyph image is owing, in the main, to the large redundancy occurring between the left and right stereoscopic images. Fig. 99 illustrates the concept of the 3D anaglyph through the splitting of the RGB channels into R and G-B. This is referred to as red-cyan.

This paper is organised as follows: Section B provides a brief discussion of some of the related. Section C discusses the method chosen for generation of the 3D anaglyph images. Section D reports the experimental results. Section E closes the paper.

B. Related Work

Dubois proposed the CIE XYZ anaglyph method which attempts to minimise the differences between the anaglyph image and the perception of the source stereoscopic image pair by minimising the Euclidean distance (L_2 norm) [183]. This is formulated as

$$\min_{V_A} \| (C \dot{\times} V) - (C_A \dot{\times} V_A) \|, \tag{45}$$

where V is a 6×1 matrix representing the normalised pixel value of the stereo pair, given by $V = \left[R_l, G_l, B_l, R_r, G_r, B_r\right]^T$; V_A is a 3×1 matrix corresponding to the pixel in the anaglyph image, given by $V = \left[R_A, G_A, B_A\right]^T$; C is a 6×6 matrix given by

$$C = \begin{bmatrix} C_s & 0 \\ 0 & C_s \end{bmatrix}, \tag{46}$$

where

$$C_s = \begin{bmatrix} 0.4641 & 0.3055 & 0.1808 \\ 0.2597 & 0.6592 & 0.0811 \\ 0.0357 & 0.1421 & 0.9109 \end{bmatrix}, \tag{47}$$

and is responsible for converting V from the RGB colour space into the CIE XYZ colour space; and C_A is a 6×3 matrix given by

$$C_A = \begin{bmatrix} A_r \\ A_l \end{bmatrix}, \tag{48}$$

with

$$A_l = \begin{bmatrix} 0.3185 & 0.0769 & 0.0109 \\ 0.1501 & 0.0767 & 0.0056 \\ 0.0007 & 0.0020 & 0.0156 \end{bmatrix} \tag{49}$$

and

$$A_r = \begin{bmatrix} 0.0174 & 0.0484 & 0.1402 \\ 0.0184 & 0.1807 & 0.0458 \\ 0.0286 & 0.0991 & 0.7662 \end{bmatrix}, \tag{50}$$

where A_l corresponds to the simulation of the red filter on the left and A_r corresponds to the simulation of the cyan filter on the right, which converts the passing lights into the CIE XYZ colour space. The symbol $\dot{\times}$ in Eq. (45) denotes matrix multiplication.

Zhang and McAllister [184] proposed using the Chebyshev distance (L_∞ norm) for the minimisation of Eq. (45). This uniform approximation method is shown to be superior in colour representation compared to the aforementioned Euclidean distance method. However, this technique is less computationally efficient owing to each pixel in the anaglyph image needing to be linearly optimised.

McAllister et al. [185] proposed the CIEL*a*b* anaglyph method, which minimises the colour distances in the CIEL*a*b* uniform colour space as opposed to the CIE XYZ colour space, as described above. In this case the CIE XYZ colour space is first determined by setting

$$C_s = \begin{bmatrix} 0.4243 & 0.3105 & 0.1657 \\ 0.2492 & 0.6419 & 0.1089 \\ 0.0265 & 0.1225 & 0.8614 \end{bmatrix}, \tag{51}$$

and the stereo pair are then converted to the L*a*b* colour space using

$$\begin{aligned} L_i &= 116 \cdot \sqrt[3]{\left(\frac{Y_i}{Y_n}\right)} - 16 \\ a_i &= 500 \cdot \left[\sqrt[3]{\left(\frac{X_i}{X_n}\right)} - \sqrt[3]{\left(\frac{Y_i}{Y_n}\right)}\right] \\ b_i &= 200 \cdot \left[\sqrt[3]{\left(\frac{Y_i}{Y_n}\right)} - \sqrt[3]{\left(\frac{Z_i}{Z_n}\right)}\right], \end{aligned} \tag{52}$$

where $i \in \{l, r\}$ and $(X_n, Y_n, Z_n) = (11.144, 100, 35.201)$, which represents the normalised tristimulus values of the white point. The matrices used for the simulation of the red and cyan filters are given by

$$A_l = \begin{bmatrix} 0.1840 & 0.0179 & 0.0048 \\ 0.0876 & 0.0118 & 0.0018 \\ 0.0005 & 0.0012 & 0.0159 \end{bmatrix} \tag{53}$$

and

$$A_r = \begin{bmatrix} 0.0153 & 0.1092 & 0.1171 \\ 0.0176 & 0.3088 & 0.0777 \\ 0.0201 & 0.1016 & 0.6546 \end{bmatrix}, \tag{54}$$

respectively.

Each anaglyph pixel is estimated by simultaneous minimisation of the Euclidean (least squares) distance of the CIEL*a*b* colours of the left eye and right eye pixels when converted from their associated RGB values. The difficulty encountered is the minimisation of the equation is ill-posed; therefore, an efficient iterative method, such as the Levenberg-Marquardt algorithm, is needed to resolve this unconstrained nonlinear problem.

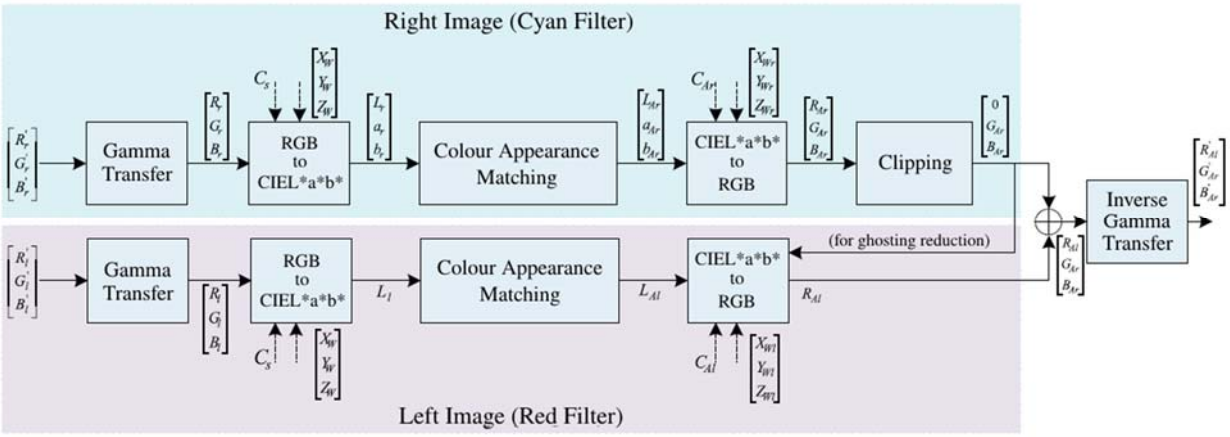


Fig. 100 Colour Appearance Attributes CIEL*a*b* Anaglyph Method [50]

The advantage of working in the CIEL*a*b* colour space, compared to the CIE XYZ colour space, is with the former the perceptual colour distance correlates well with the Euclidean distance; in addition, the CIEL*a*b* colour space more precisely simulates the perception of the HVS owing to the incorporation of nonlinear response compression and chromatic adaptation transform algorithms.

Li et al. [50] proposed that the matching of colour appearance attributes (CAA), such as lightness, saturation and hue, be considered, in addition to the XYZ or L*a*b* values. The key benefit of working within the CIEL*a*b* colour space, as opposed to the CIE XYZ colour space, is the former allows for a more direct method of translating these aforementioned colour appearance attributes. Fig. 100 provides an illustration of the proposed CAA CIEL*a*b* anaglyph method.

Gamma and inverse gamma correction is implemented to compensate for the nonlinear transformation of the captured RGB values in the reproduction of the colour component for the LCD display. The clipping subcomponent is responsible for keeping each channel range between 0 and 1. An advantage of using the CAA CIEL*a*b* anaglyph method is the minimisation algorithm is in a closed form solution, as opposed to the ill-posed iterative optimisation technique proposed in the CIEL*a*b* anaglyph method. Furthermore, the proposed CAA CIEL*a*b* anaglyph method has shown to noticeably improve on the effects of the three most pronounced shortcomings generally encountered with anaglyphs, viz. colour distortions (also referred to as chrominance accuracy), retinal rivalry and ghosting effects.

C. Proposed Approach

The primary focus of this research is 2D-to-3D conversion, which involves the synthesis of disparity image pairs from single 2D LDOF images. This essentially concluded at the end of the hole-filling sub-process within the DIBR process discussed in the previous section *Stereoscopic Image Synthesis of Single 2D Low Depth-of-Field Images using Depth Image-Based Rendering* on *p*. 134. The mode of display is only aesthetic and not an aspect of the 2D-to-3D conversion

process per se. Therefore, emphasis is placed on employing anaglyph as a simple, efficient and cost effective means of verifying the subjective efficacy of the proposed autonomous 2D-to-3D conversion model and not on the development of a newer or better method of anaglyph image generation or presentation. In the proposed model the anaglyph image is generated from the synthesised left and right disparity images using the method proposed by Li et al. [50].

D. Results

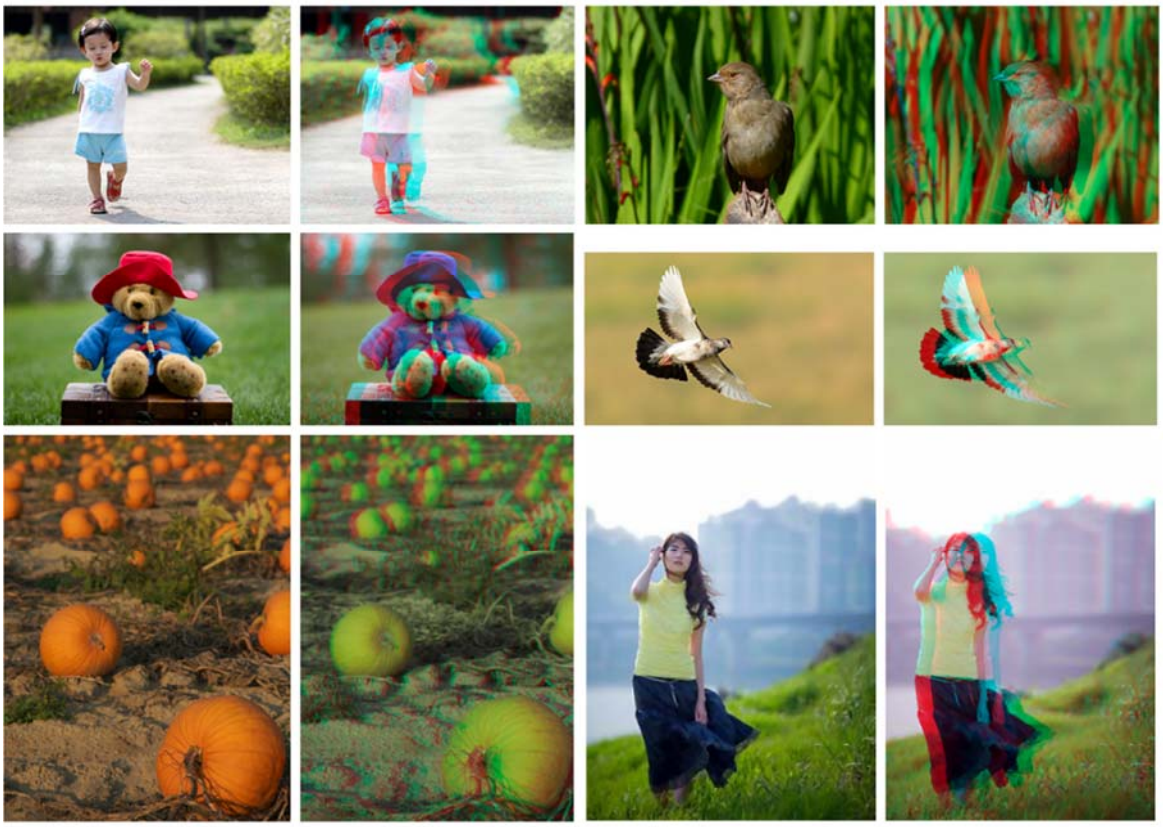


Fig. 101 3D images of training data. The original images together with their 3D anaglyph interpretations based on the synthesised left and right disparity images.

Some of the representative results obtained for the training data are illustrated in Fig. 101. Owing to space limitations only a limited number of results of the total dataset of compressed test images are presented in this section. The appendix on *p*. 188 contains a selection of uncompressed or in some instances, optimally resized, training and test results. The complete database of training and test LDOF images and their associated anaglyphs is available from the author at serenr@gmail.com.

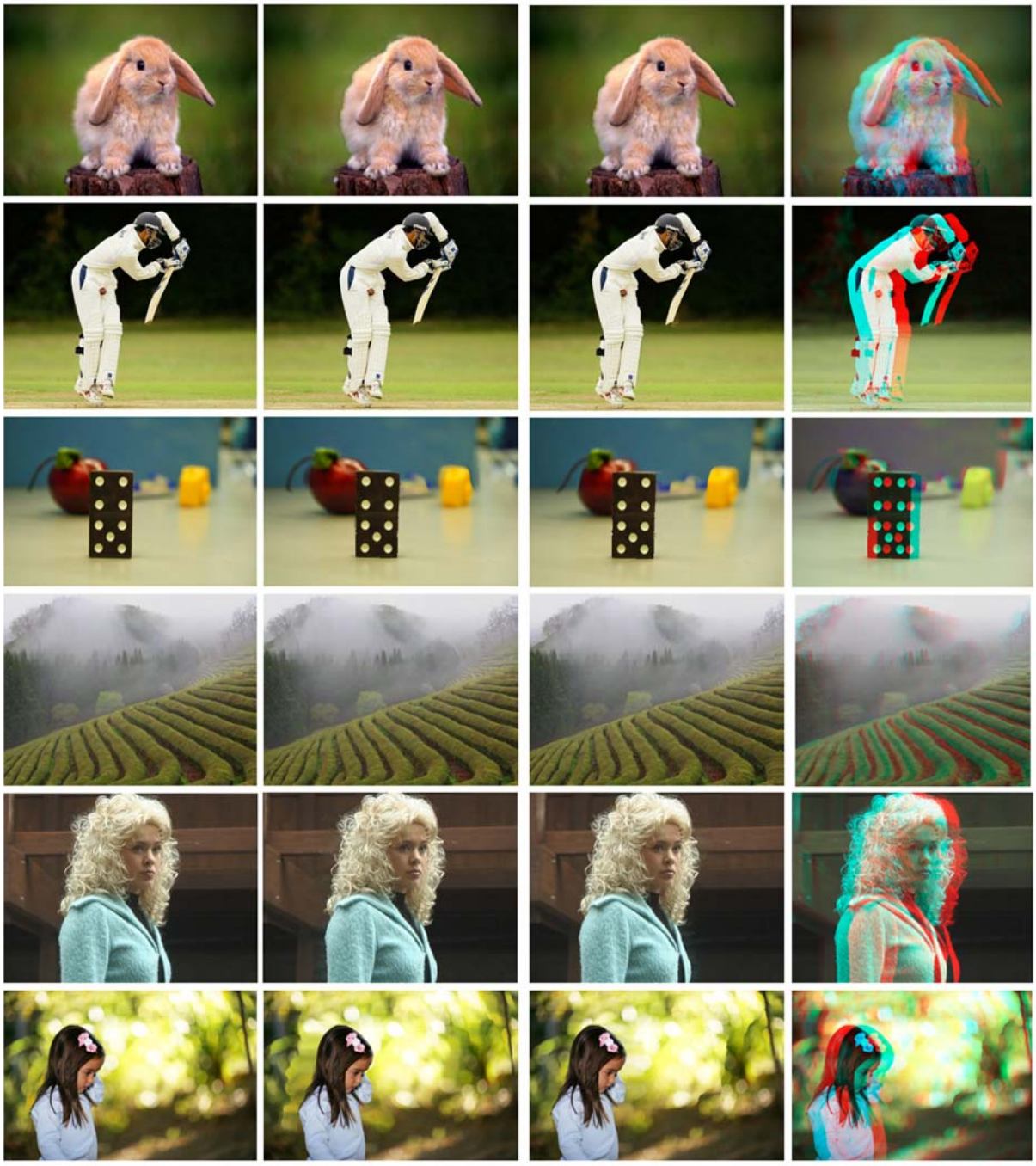


Fig. 102 3D images of test data – Set 3. The original images, synthesised left and right disparity images and associated 3D anaglyph interpretations are presented.

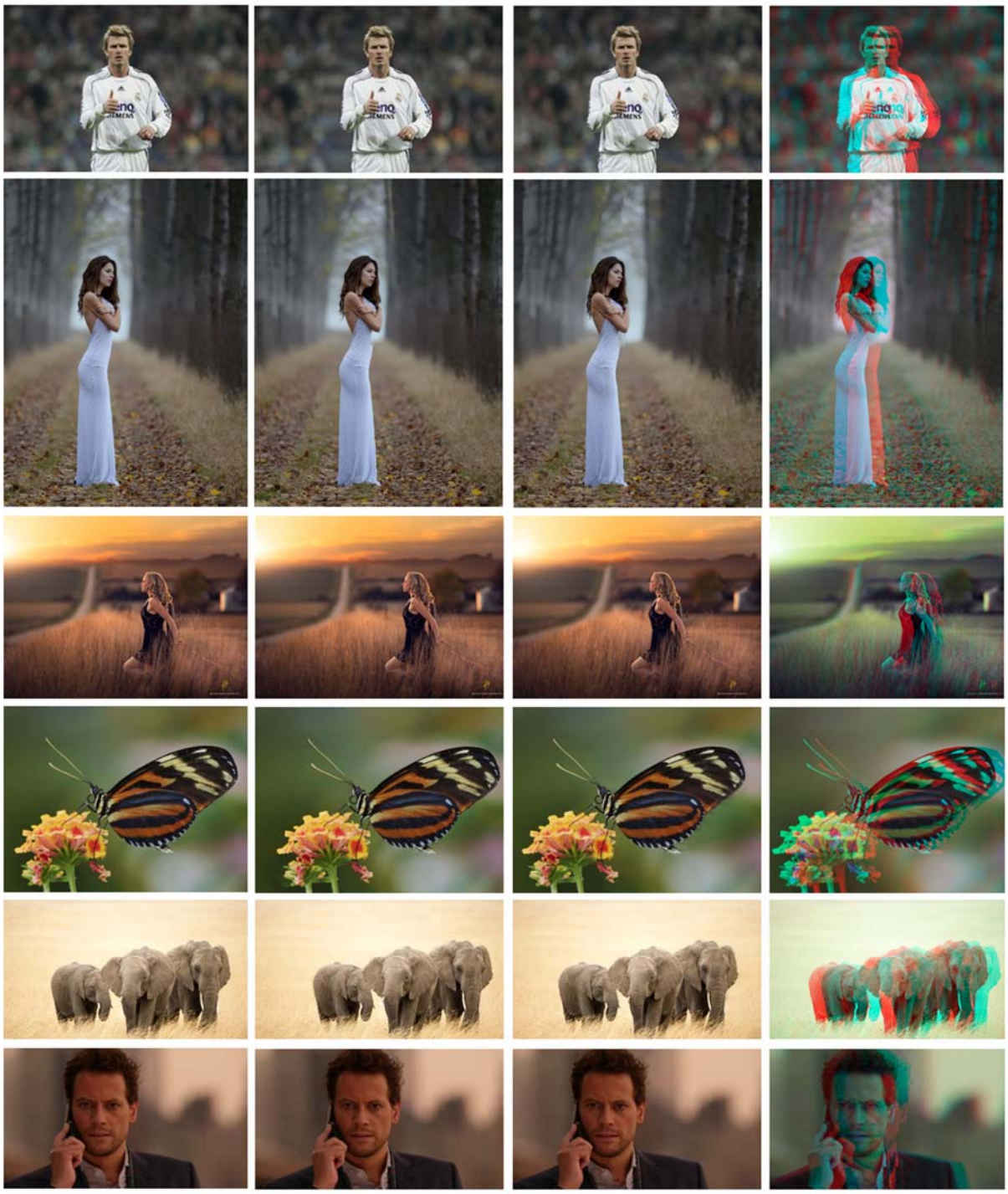


Fig. 103 3D images of test data – Set 1. The original images, synthesised left and right disparity images and associated 3D anaglyph interpretations are presented.

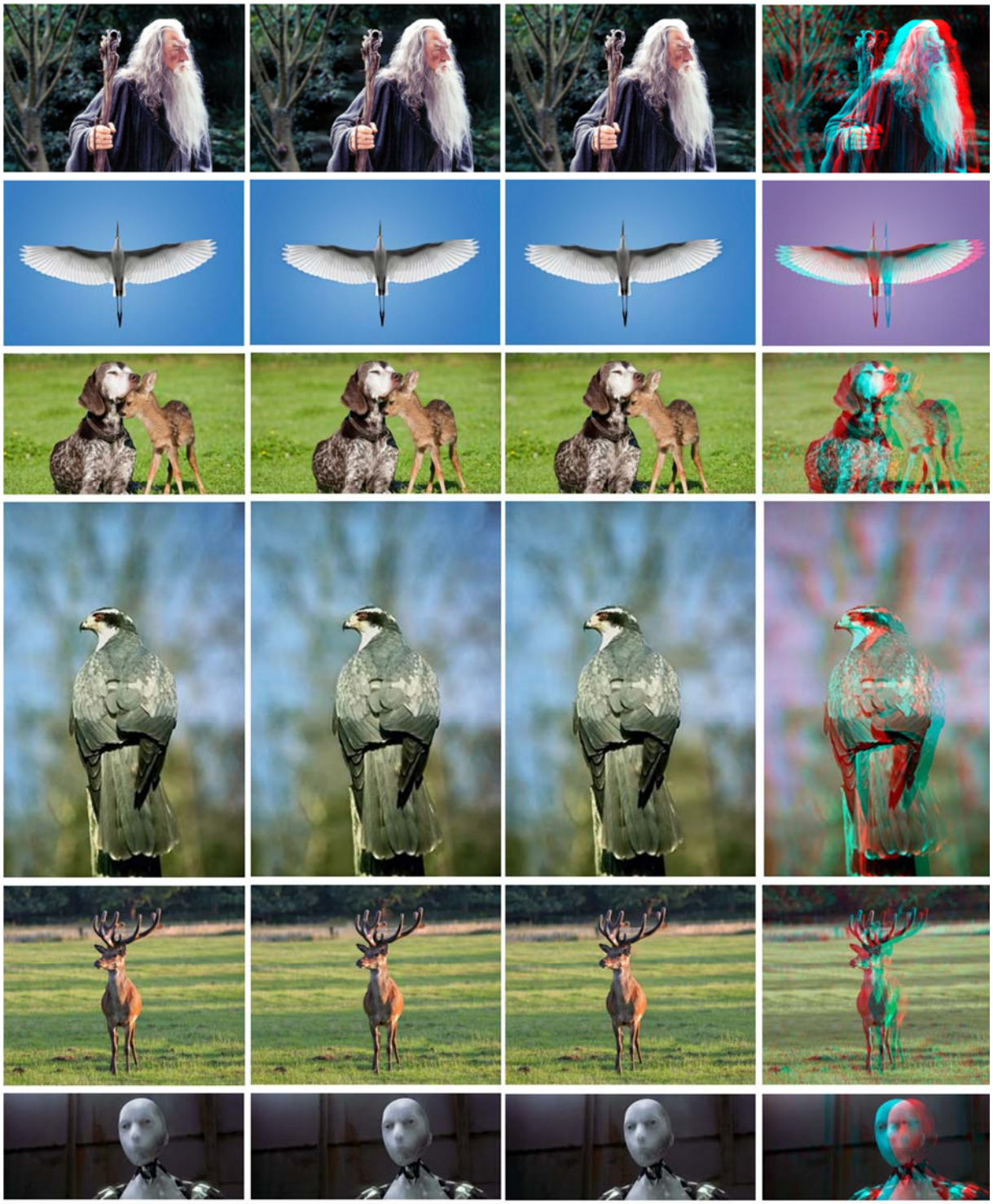


Fig. 104 3D images of test data – Set 2. The original images, synthesised left and right disparity images and associated 3D anaglyph interpretations are presented.

E. Conclusion

The method proposed by Li et al. [50] is considered for the unsupervised generation of 3D anaglyph image from two disparity images. Although anaglyph is qualitatively inferior, when compared to the other two aforementioned stereoscopic visualisation methods this technique is the only one presentable in standard colour print format. Moreover, the 3D experience is achievable using inexpensive colour filtered glasses and ordinary computer monitors, without the need for any additional 3D compatible hardware and/or software.

VIII. CONCLUSION

A. Summary

This research presents a novel method for the automatic synthesis of stereoscopic image pairs from single 2D low depth-of-field (LDOF) images. Results show that this is achievable through the unsupervised generation of depth maps and the subsequent extrapolation of disparity image pairs through the use of depth image-based rendering (DIBR).

Results show that the accuracy of the synthesised stereoscopic images is inherently dependent on the accuracy of the depth map. In turn the accuracy of the depth map is inherently dependent on the accuracy of the segmentation of the objects and regions in the image.

In LDOF images the emphasis is usually placed on a distinct object-of-interest (OOI) in the scene. A LDOF image therefore may be described as being constituted of two regions. These include the OOI and the non-OOI region. In the proposed model segmentation is performed in three stages. Firstly, the region-of-interest (ROI) is extracted and secondly, the object-of-interest (OOI) is delineated within the ROI and thirdly, the non-OOI region is sub-segmented.

Results show that the interrogation and correlation of edge, gradient and higher-order statistics (HOS) saliencies together with *k*-means segmentation provides an extremely robust automatic method for the extraction of the ROI. Moreover, the results show that the proposed method outperforms several of the previously proposed state-of-the-art approaches [41, 42, 51-54].

Sub-segmentation of the ROI is only necessary for scenarios where the ROI contains both an in focus OOI and an in focus non-OOI region. It is therefore necessary to discern between the different ROI scenarios. In other words, a classification of the LDOF image is required. The challenge is that owing to there being no a priori reference for the OOI the problem becomes ill-posed. However, the results show that the interrogation of the dominant gradients within the ROI is an extremely effective means of identifying the OOI and also providing a rudimentary or baseline reference for the approximation of the OOI. Subsequently, an adequate approximation of the OOI may be extrapolated through the expansion and refinement of the baseline OOI reference using edge saliencies and *k*-means segmentation. Results show that this approximation of the OOI together with the ROI provides an extremely accurate means of classifying the LDOF image.

Moreover, results show that an accurate delineation of the non-OOI region within the ROI is possible using *k*-means clustering together with the approximation of the OOI as a reference. The subsequent removal of the non-OOI region from the ROI results in the matting of the closed boundary region of the OOI.

The relationship between the OOI and the entire (focussed and defocussed) non-OOI region in a LDOF may be described as either being equidistant-based or gradient-plane-based. For the former the entire non-OOI region may be described as being relatively equidistant behind the

OOI. For the latter the objects and regions within the non-OOI region may be described as being either in front of or behind the OOI at different relative depths.

Sub-segmentation of the non-OOI region is only necessary for the gradient-plane-based scenario. It is therefore necessary to distinguish between the two non-OOI scenarios. Results show that by correlating the *k*-means segmentation with the binary quantisation segmentation of the non-OOI region it is possible, with a high degree of accuracy, to determine if the non-OOI region is equidistant-based or gradient-plane-based.

In the equidistant-based scenario the relative depths of the non-OOI region are assigned using depth from defocus (DfD). The DfD method chosen in the proposed model is the Gaussian re-blur technique.

In the gradient-plane-based scenario the non-OOI region is sub-segmented using *k*-means and binary quantisation segmentation. Correlation of the clusters using Gestalt-based principles is used to delineate and subsequently label the objects and regions in the non-OOI region. The assigning of relative depths to the segmented objects (including the OOI) and regions is achieved using a gradient-plane. The incremental relative depth of the gradient-plane is determined by correlating the Gestalt-based ground region with vanishing point (VP) detection.

VP detection is achieved through the analysis of the intersecting dominant straight edges in the image directly in the image-plane parametric space. The results show that the proposed unsupervised VP detection method either matches or outperforms several of the previously proposed approaches [23, 56-59].

Results show that some of inconsistencies in the depth, such as blocky artefacts and outliers, are alleviated through the application of a bilateral edge-preserving filter. The proposed automatic depth map generation method outperforms several of the previously proposed approaches [60-65].

The stereoscopic image pairs are synthesised using the depth map together with depth image-based rendering (DIBR). The results show that some of the dis-occlusions are either negated or alleviated by applying an asymmetrical bilateral edge-preserving filter prior to the 3D warping process. The remaining dis-occlusions are resolved through horizontal interpolation that takes into account the direction of warp. To alleviate some of the stripy artefacts Gaussian and circular averaging filtering is applied to the newly interpolated regions followed by average filtering on the border transitions. This proposed staged filtering approach effectively reduces the discontinuities while preserving the subjective view quality. Results show that the proposed method for the unsupervised resolution of the dis-occlusions generated during the automatic 2D-to-3D conversion of single 2D LDOF images outperforms other previously proposed approaches [62, 65].

Although computational efficiency is taken into account in this research, accuracy and quality are given prominence over the need for real-time speed. This research shows that the automatic 2D-to-3D conversion of single 2D low depth-of-field images is achievable through firstly, the delineation and labelling of the objects and regions in the image through the interrogation of edge, gradient and higher-order statistics together with *k*-means and binary

quantisation segmentation as well as the application of Gestalt principles and secondly, the relative assignment of depth the objects and regions through Gaussian re-blur analysis and vanishing point detection and thirdly, the synthesis of the stereoscopic image pair using depth image-based rendering.

B. Future Research

Some of the future work may include:

1. The expansion of the entire model to incorporate monocular LDOF sequential video frames as well as real-time functionality.
2. The use of more scenario-specific segmentation methods as well as the possible incorporation of machine learning to more effectively identify and subsequently assign relative depth values to the objects and regions in an image.
3. Improved vanishing point detection to account for more natural environments as well as investigate unsupervised techniques for the use in the classification of natural and artificial environments.
4. More precise hole-filling techniques using machine learning and Bayesian inference as well as investigate autonomous techniques to account for sequential video.
5. The possible incorporation of high DOF sequential video.

IX. REFERENCES

[1] I. Sexton and P. Surman, "Stereoscopic and Autostereoscopic Display Systems: An In-Depth Review of Past, Present, and Future Technologies," *IEEE Signal Processing Magazine,* vol. 16, pp. 85–99, 1999. doi: 10.1109/79.768575

[2] M. Seymour. (2012) Art of Stereo Conversion: 2D to 3D. *fxguide*.

[3] R. Zone, *Stereoscopic Cinema and the Origins of 3-D Film, 1838-1952*. Lexington, Kentucky: The University Press of Kentucky, 2007. ISBN: 978-0813124612

[4] R. Zone, *3-D fimmakers: Conversations with Creators of Stereoscopic Motion Pictures* vol. 119: Scarecrow Press, 2005. ISBN: 978-0810854376

[5] PwC, "Waiting for the Next Wave: 3D Entertainment 2012," Price Waterhouse Coopers International Limited, 2012. Available: http://www.pwc.com/gx/en/entertainment-media/publications/waiting-for-the-next-wave-3d-entertainment-2012.jhtml

[6] M. Murphy, P. Jean, and J. Myint, "Inside the 3-D Conversion of 'Titanic'," in *The New York Times*, ed, 2012.

[7] I. Failes. (2013) Welcome (back) to Jurassic Park. *fxguide*. Available: http://www.fxguide.com/featured/welcome-back-to-jurassic-park/

[8] J. P. Frisby, *Seeing: Illusion, Brain and Mind*. Oxford: Oxford University Press, 1979.

[9] R. L. Gregory, *Eye and Brain: The Psychology of Seeing*: Princeton University Press, 1998.

[10] J. A. Norling, "The Stereoscopic Art: A Reprint," *Journal of the Society of Motion Picture and Television Engineers,* vol. 60, pp. 268-308, 1953.

[11] W. Richards, "Configuration Stereopsis: A New Look at the Depth–Disparity Relation," *Spatial Vision,* vol. 22, pp. 91-103, 2009.

[12] W. Richards, "Visual Space Perception," in *Handbook of Perception*, E. C. Carterette, Friedman, M.P., Ed., ed: Academic Press, 1975, pp. 351-386.

[13] C. H. Graham, "Visual Space Perception," in *Vision and Visual Perception*, C. H. Graham, Ed., ed New York: Wiley, 1965.

[14] L. Lipton, *Foundations of the Stereoscopic Cinema*. New York: Van Nostrand Reinhold Company, 1982.

[15] L. Lipton. (1997). *The Stereographics® Developer's Handbook - Background on Creating Images for CrystalEyes® and SimulEyes®*: Stereographics Corporation, 1997. Available: www.cs.unc.edu/Research/stc/FAQs/Stereo/stereo-handbook.pdf

[16] C. Wheatstone, "On some remarkable, and hitherto unobserved, phenomena of binocular vision (Part the first)," *Royal Society of London, Transactions of the,* pp. 371-394, 1838.

[17] F. Devernay and P. Beardsley, "Stereoscopic Cinema," in *Image and Geometry Processing for 3-D Cinematography*, R. Ronfard, Taubin, G., Ed., ed Berlin Heidelberg: Springer, 2010.

[18] M. Merzenich. (2014, 23 June). *Brain Memory, How Vision Works, Human Visual System*. San Francisco, CA: Posit Science, 2014. Available: http://www.brainhq.com/brain-resources/brain-facts-myths/how-vision-works

[19] L. Zhang, C. Vázquez, and S. Knorr, "3D-TV Content Creation: Automatic 2D-to-3D Video Conversion," *Broadcasting, IEEE Transactions on,* vol. 57, pp. 372-383, 2011. doi: 10.1109/TBC.2011.2122930

[20] B. Rogers and M. Graham, "Motion Parallax as an Independent Cue for Depth Perception," *Perception,* vol. 8, pp. 125-134, 1979. doi: 10.1068/p080125

[21] B. Brillaut-O'Mahony, "New Method for Vanishing Point Detection," *Graphical Models and Image Processing (CVGIP): Image Understanding,* vol. 54, pp. 289–300, 1991. doi: 10.1016/1049-9660(91)90069-2

[22] J. Kogecka and W. Zhang, "Efficient Computation of Vanishing Points," in *IEEE International Conference on Robotics and Automation (ICRA), Proceedings of the*, 2002, pp. 223-228. doi: 10.1109/ROBOT.2002.1013365

[23] S. Battiato, A. Capra, S. Curti, and M. La Cascia, "3D Stereoscopic Image Pairs by Depth-map Generation," in *3D Data Processing, Visualization and Transmission (3DPVT), Proceedings of the 2nd International Symposium on*, 2004, pp. 124-131. doi: 10.1109/TDPVT.2004.1335185

[24] K. R. Thórisson, "Simulated Perceptual Grouping: An Application to Human-Computer Interaction," in *Cognitive Science Society, Proceedings of the Sixteenth Annual Conference of the*, Atlanta, Georgia, 1994, pp. 876-881. doi: 10.1.1.212.4867

[25] E. Trucco, *Introductory Techniques for 3-D Computer Vision*: Prentice Hall, 1998. ISBN: 978-0132611084

[26] A. P. Pentland, "Local Shading Analysis," *Pattern Analysis and Machine Intelligence,* vol. 6, pp. 170-187, 1984. doi: 10.1109/TPAMI.1984.4767501

[27] B. K. P. Horn, "Obtaining Shape from Shading Information," in *The Psychology of Computer Vision*, P. H. Winston, Ed., ed New York: McGraw-Hill, 1975, pp. 115-155.

[28] M. Shah, *Fundamentals of Computer Vision*. Orlando, FL: University of Central Florida, 1997.

[29] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant Dehazing of Images using Polarization," in *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, 2001, pp. 325-332. doi: 10.1109/CVPR.2001.990493

[30] D. A. Forsyth, "Shape from Texture and Integrability," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* Vancouver, BC, 2001, pp. 447–452. doi: 10.1109/ICCV.2001.937659

[31] A. M. Loh and R. Hartley, "Shape from Non-Homogeneous, Non-Stationary, Anisotropic, Perspective Texture," in *British Machine Vision Conference, Proceedings of the*, Oxford, U.K., 2005, pp. 8.1-8.10. doi: 10.5244/C.19.8

[32] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Müller, *et al.*, "Three-Dimensional Video Postproduction and Processing," *Proceedings of the IEEE,* vol. 99, pp. 607-625, 2011. doi: 10.1109/JPROC.2010.2098350

[33] H. Murata, T. Okino, T. Iinuma, S. Yamashita, S. Tanase, K. Terada, *et al.*, "Conversion of Two-Dimensional Image to Three Dimensions," in *Society of Information Displays (SID) Symposium Digest of Technical Papers*, 1995, pp. 859-862.

[34] T. Okino, H. Murata, K. Taima, T. Iinuma, and K. Oketani, "New Television With 2D/3D Image Conversion Technologies," in *Stereoscopic Displays and Virtual Reality Systems III (SPIE 2563), Proceedings of*, San Jose, CA, 1996, pp. 96-103. doi: 10.1117/12.237421

[35] G. Zhang, W. Hua, X. Qin, T.-T. Wong, and H. Bao, "Stereoscopic Video Synthesis from a Monocular Video," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 13, pp. 686-696, 2007. doi: 10.1109/TVCG.2007.1032

[36] C. Fehn, P. Kauff, M. Op De Beeck, F. Ernst, W. IJsselsteijn, M. Pollefeys, *et al.*, "An Evolutionary and Optimised Approach on 3D-TV," in *International Broadcast Conference, Proceedings of the*, 2002, pp. 357-365. Available: http://www.cs.unc.edu/~marc/pubs/FehnIBC02.pdf

[37] S.-Y. Kim, S.-B. Lee, and Y.-S. Ho, "Three-Dimensional Natural Video System Based on Layered Representation of Depth Maps," *Consumer Electronics, IEEE Transactions on,* vol. 52, pp. 1035-1042, 2006. doi: 10.1109/TCE.2006.1706504

[38] X. Cao, A. C. Bovik, Y. Wang, and Q. Dai. (2011) Converting 2D Video to 3D: An Efficient Path to a 3D Experience. *MultiMedia, IEEE*. 12-17.

[39] C. Fehn, "A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR)," in *International Conference on Visualization, Imaging, and Image Processing, Proceedings of the 3rd International Association of Science and Technology for Development (IASTED)*, Benalmádena, Spain, 2003, pp. 482-487. Available: http://www.actapress.com/Content_Of_Proceeding.aspx?ProceedingID=231

[40] We-Are-Instrument. (2013). Bike Lift [Online]. Available: http://weareinstrument.com

[41] C. Kim, "Segmenting a Low-Depth-of-Field Image Using Morphological Filters and Region Merging," *Image Processing, IEEE Transactions on,* vol. 14, pp. 1503-1511, 2005. doi: 10.1109/TIP.2005.846030

[42] H. Li and K. N. Ngan, "Unsupervized Video Segmentation With Low Depth of Field," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 17, pp. 1742-1751, 2007. doi: 10.1109/TCSVT.2007.903326

[43] H. D. Cheng, X. H. Jiang, Y. Sun, and J. Z. Wang, "Color Image Segmentation: Advances and Prospects," *Pattern Recognition,* vol. 34, pp. 2259-2281, 2001. doi: 10.1016/S0031-3203(00)00149-7

[44] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A Bayesian Approach to Digital Matting," in *IEEE CVPR, Proceedings of the*, 2001, pp. 264–271.

[45] A. Levin, D. Lischinski, and Y. Weiss, "A Closed-Form Solution to Natural Image Matting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 30, 2008. doi: 10.1109/TPAMI.2007.1177

[46] P. Debevec, C. Taylor, and J. Malik, "Modeling and Rendering Architecture from Photographs: A hybrid Geometry- and Image-based Approach," in *Computer Graphics and Interactive Techniques, Proceedings of the SIGGRAPH 23rd Annual Conference on* New York, NY, USA, 1996, pp. 11-20. doi: 10.1145/237170.237191

[47] L. Zhang and W. J. Tam, "Stereoscopic Image Generation based on Depth Images for 3DTV," *Broadcasting, IEEE Transactions on,* vol. 51, pp. 191-199, 2005.

[48] Y. K. Park, K. Jung, Y. Oh, S. Lee, J. K. Kim, G. Lee*, et al.*, "Depth-Image-Based Rendering for 3DTV Service Over T-DMB," *Signal Processing: Image Communication,* vol. 24, pp. 122-136, 2009. doi: 10.1016/j.image.2008.10.008

[49] C. Vázquez, W. J. Tam, and F. Speranza, "Stereoscopic Imaging: Filling Disoccluded Areas in Depth Image-Based Rendering," in *Three-Dimensional TV, Video, and Display V, Proceedings of SPIE*, Boston, MA, 2006. doi: 10.1117/12.685047

[50] S. Li, L. Ma, and K. N. Ngan, "Anaglyph Image Generation by Matching Color Appearance Attributes," *Signal Processing: Image Communication,* vol. 28, pp. 597-607, 2013. doi: 10.1016/j.image.2013.03.004

[51] T. Chen and H. Li, "Segmenting Focused Objects based on the Amplitude Decomposition Model," *Pattern Recognition Letters,* vol. 33, pp. 536–1542, 2012. doi: 10.1016/j.patrec.2012.04.014

[52] H. Li and K. N. Ngan, "Learning to Extract Focused Objects From Low DOF Images," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 21, pp. 1571-1580, 2011.

[53] Z. Liu, W. Li, L. Shen, Z. Han, and Z. Zhang, "Automatic Segmentation of Focused Objects from Images with Low Depth of Field," *Pattern Recognition Letters,* vol. 31, pp. 572-581, 2010. doi: 10.1016/j.patrec.2009.11.016

[54] Z. Ye and C.-C. Lu, "Unsupervised Multiscale Focused Objects Detection using Hidden Markov Tree," in *Computer Vision, Pattern Recognition and Image Processing, Proceedings of the Int. Conf. on*, Durham, NC, 2002, pp. 812-815.

[55] G. Rafiee, "Automatic Region-of-Interest Extraction in Low Depth-of-Field Images," Doctor of Philosophy, School of Electrical And Electronic Engineering, Newcastle University, United Kingdom, 2013.

[56] S. Battiato, S. Curti, M. La Cascia, M. Tortora, and E. Scordato, "Depth-Map Generation by Image Classification," in *SPIE, Proceedings of the*, 2004, pp. 95-104. doi: 10.1117/12.526634

[57] E. J. Chappero, R. A. Guerrero, and F. J. Serón, "Vanishing Point Estimation from Monocular Images," in *9th International Information and Telecommunication Technologies Symposium (I2TS), Proceedings of the*, 2010, pp. 177-182. Available: http://www.mendeley.com/download/public/15725273/4937673802/77139b5789d10be239bda2d03831e8c4bba16903/dl.pdf

[58] P. Denis, J. H. Elder, and F. Estrada, "Efficient Edge-Based Methods for Estimating Manhattan Frames in Urban Imagery," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 197-210. doi: 10.1007/978-3-540-88688-4_15

[59] F. Schaffalitzky and A. Zisserman, "Planar Grouping for Automatic Detection of Vanishing Lines and Points," *Image and Vision Computing,* vol. 18, pp. 647-658, 2000. doi: 10.1016/S0262-8856(99)00069-4

[60] S. Bae and F. Durand, "Defocus Magnification," in *Eurographics 2007, Proceedings of*, Prague, Czech Republic, 2007. doi: 10.1111/j.1467-8659.2007.01080.x

[61] G. Guo, N. Zhang, L. Huo, and W. Gao, "2D to 3D Convertion (sic) Based on Edge Defocus and Segmentation," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on* Las Vegas, NV, 2008, pp. 2181-2184. doi: 10.1109/ICASSP.2008.4518076

[62] S. A. Valencia and R. M. Rodriguez-Dagnino, "Synthesizing Stereo 3D Views from Focus Cues in Monoscopic 2D Images " in *Stereoscopic Displays and Virtual Reality Systems X, SPIE Proceedings of*, Santa Clara, CA, 2003. doi: 10.1117/12.474113

[63] S. Zhuo and T. Sim, "Defocus Map Estimation from a Single Image," *Pattern Recognition,* vol. 44, pp. 1852-1858, 2011. doi: 10.1016/j.patcog.2011.03.009

[64] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar, "Estimating Spatially Varying Defocus Blur From A Single Image," *IEEE Transactions on Image Processing,* vol. 22, pp. 4879-4891, 2013. doi: 10.1109/tip.2013.2279316

[65] J. Ko, M. Kim, and C. Kim, "2D-To-3D Stereoscopic Conversion: Depth-Map Estimation in a 2D Single-View Image," in *Applications of Digital Image Processing XXX, SPIE Proceedings of*, 2007, pp. 1-9. doi: 10.1117/12.736131

[66] Easy Basic Photography, "Deep Depth of Field vs Shallow Depth of Field", 2012, Available: http://easybasicphotography.com/

[67] C. Kim, J. Park, J. Lee, and J.-N. Hwang, "Fast Extraction of Objects of Interest from Images with Low Depth of Field," *ETRI Journal,* vol. 29, pp. 353-362, 2007. doi: 10.4218/etrij.07.0106.0173

[68] C. S. Won, K. Pyun, and R. M. Gray, "Automatic Object Segmentation in Images with Low Depth of Field," in *Image Processing, Proceedings of International Conference on*, 2002. doi: 10.1109/ICIP.2002.1039094

[69] D.-M. Tsai and H.-J. Wang, "Segmenting Focused Objects in Complex Visual Images," *Pattern Recognition Letters,* vol. 19, pp. 929-940, 1998. doi: 10.1016/S0167-8655(98)00078-6

[70] J. Z. Wang, J. Li, R. M. Gray, and G. Wiederhold, "Unsupervised Multiresolution Segmentation for Images with Low Depth of Field," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 23, pp. 85-90, 2001. doi: 10.1109/34.899949

[71] J. Hartigan and M. Wong, "Algorithm AS136: A K-means Clustering Algorithm," *Applied Statistics,* vol. 28, pp. 100–108, 1979. doi: 10.2307/2346830

[72] K. Fukunaga and L. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *Information Theory, IEEE Transactions on,* vol. 21, pp. 32-40, 1975. doi: 10.1109/TIT.1975.1055330

[73] M. Ben Salah, A. Mitiche, and I. Ben Ayed, "Multiregion Image Segmentation by Parametric Kernel Graph Cuts," *IEEE Transactions on Image Processing,* vol. 2, pp. 545-557, 2011. doi: 10.1109/TIP.2010.2066982

[74] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 603-619, 2002. doi: 10.1109/34.1000236

[75] Y. Boykov and M.-P. Jolly, "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images," in *International Conference on Computer Vision, Proceedings of the*, Vancouver, Canada, 2001, pp. 105-112.

[76] Y. Boykov and G. Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation," *International Journal of Computer Vision,* vol. 60, pp. 109-131, 2006. doi: 10.1007/s11263-006-7934-5

[77] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *International Journal of Computer Vision,* vol. 1, pp. 321-331, 1988. doi: 10.1007/BF00133570

[78] T. F. Chan and L. A. Vese, "Active Contours Without Edges," *IEEE Transactions on Image Processing,* vol. 10, pp. 266-277, 2001. doi: 10.1109/83.902291

[79] J. Sourati, D. Erdogmus, J. G. Dy, and D. H. Brooks, "Accelerated Learning-Based Interactive Image Segmentation Using Pairwise Constraints," *IEEE Transactions on Image Processing,* vol. 23, pp. 3057-3070, 2014. doi: 10.1109/TIP.2014.2325783

[80] B.-B. Chai, J. Vass, and X. Zhuang, "Significance-Linked Connected Component Analysis for Wavelet Image Coding," *Image Processing, IEEE Transactions on,* vol. 8, pp. 774-784, 1999. doi: 10.1109/83.766856

[81] F. Rooms, A. Pizurica, and W. Philips, "Estimating Image Blur in the Wavelet Domain," in *Computer Vision, Proceedings of the 5th Asian Conference on*, 2002, pp. 210-215.

[82] V. P. Namboodiri and S. Chaudhuri, "Recovery of Relative Depth from a Single Observation using an Uncalibrated (Real-Aperture) Camera," in *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, 2008, pp. 1-6. doi: 10.1109/CVPR.2008.4587779

[83] Y.-W. Tai, H. Tang, M. S. Brown, and S. Lin, "Detail Recovery for Single-image Defocus Blur," *Computer Vision and Applications, IPSJ Transactions on,* vol. 1, pp. 1-10, 2009.

[84] M. McGuire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand, "Defocus Video Matting," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH,* vol. 24, pp. 567-576, 2005. doi: 10.1145/1073204.1073231

[85] J. Park and C. Kim, "Performance Improvement of Object-of-Interest Extraction from the Low Depth-of-Field Image Using Color-Based HOS (Higher-Order Statistics)," in *Korean Signal Processing Conference (KSPC), Proceedings of the*, 2005, p. 109.

[86] S. Mallat, "Wavelets for a Vision," *Proceedings of the IEEE,* vol. 84, pp. 604-614, 1996. doi: 10.1109/5.488702

[87] A. P. Pentland, "A New Sense for Depth of Field," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 9, pp. 523-531, 1987. doi: 10.1109/TPAMI.1987.4767940

[88] R. M. Haralick and L. G. Shapiro, "Image Segmentation Techniques," *Computer Vision, Graphics and Image Processing,* vol. 29, pp. 100-132, 1985. doi: 10.1117/12.948400

[89] T. Pavlidis, "Image Analysis," *Annual Review of Computer Science,* vol. 3, pp. 121-146, 1988. doi: 10.1146/annurev.cs.03.060188.001005

[90] N. R. Pal and S. K. Pal, "A Review on Image Segmentation Techniques," *Pattern Recognition,* vol. 26, pp. 1274-1294, 1993. doi: 10.1016/0031-3203(93)90135-j

[91] Y. J. Zhang, "Evaluation and Comparison of Different Segmentation Algorithms," *Pattern Recognition Letters,* vol. 18, pp. 963-974, 1997. doi: 10.1016/S0167-8655(97)00083-4

[92] W.-Z. Kang, Q.-Q. Yang, and R.-P. Liang, "The Comparative Research on Image Segmentation Algorithms," in *Education Technology and Computer Science (ETCS), Proceedings of the 1st International Workshop on*, Wuhan, Hubei, 2009, pp. 703-707. doi: 10.1109/ETCS.2009.417

[93] K. J. and I. J., "On Threshold Selection using Clustering Criteria," *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 15, pp. 652-655, 1985. doi: 10.1109/TSMC.1985.6313443

[94] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3 ed. New Jersey: Pearson Education, Inc., 2008.

[95] J. Canny, "A Computational Approach to Edge Detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 8, pp. 679-698, 1986. doi: 10.1109/TPAMI.1986.4767851

[96] P. L. Palmer, H. Dabis, and J. Kittler, "A Performance Measure for Boundary Detection Algorithms," *Computer Vision and Image Understanding,* vol. 63, pp. 476–494, 1996. doi: 10.1006/cviu.1996.0036

[97] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 9, pp. 62-66, 1979. doi: 10.1109/TSMC.1979.4310076

[98] K. C. Strasters and J. J. Gerbrands, "Three-Dimensional Image Segmentation using a Split, Merge and Group Approach," *Pattern Recognition Letters,* vol. 12, pp. 307-325, 1991. doi: 10.1016/0167-8655(91)90414-H

[99] C. R. Jung, "Multiscale Image Segmentation using Wavelets and Watersheds," in *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI) XVI, Proceedings of the*, 2003, pp. 278–284. doi: 10.1109/SIBGRA.2003.1241020

[100] H. D. Cheng and Y. Sun, "A Hierarchical Approach to Color Image Segmentation Using Homogeneity," *Image Processing, IEEE Transactions on,* vol. 9, pp. 2071-2082, 2000. doi: 10.1109/83.887975

[101] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref, "Automatic Image Segmentation by Integrating Color-Edge Extraction and Seeded Region Growing," *IEEE Transactions on Image Processing,* vol. 10, pp. 1454-1466, 2001. doi: 10.1109/83.951532

[102] J. Macqueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Mathematical Statistics and Probability, Proceedings of the 5th Berkeley Symposium on*, 1967, pp. 281-297.

[103] A. K. Jain, M. N. Murty, and P. J. Flynn "Data Clustering: A Review," *ACM Computing Surveys (CSUR),* vol. 31, pp. 264-323, 1999. doi: 10.1145/331499.331504

[104] M. T. Orchard and C. A. Bouman, "Color Quantization of Images," *Signal Processing, IEEE Transactions on,* vol. 39, pp. 2677-2690, 1991. doi: 10.1109/78.107417

[105] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*: Society for Industrial and Applied Mathematics (SIAM), 2007. ISBN: 978-0898716238

[106] M. R. Luo, G. Cui, and B. Rigg, "The Development of the Colour-Difference Formula: CIEDE2000," *Color Research and Application,* vol. 26, pp. 340-350, 2001.

[107] CIE, "Improvement to Industrial Colour-Difference Evaluation," Central Bureau of the CIE, Vienna, Rep. 142-2001, 2001. Available: http://div1.cie.co.at/?i_ca_id=551&pubid=42

[108] C.-H. Chou, K.-C. Liu, and C.-S. Lin, "Perceptually Optimized JPEG2000 Coder Based On CIEDE2000 Color Difference Equation," presented at the International Conference on Image Processing (ICIP), Proceedings of the IEEE, 2005. doi: 10.1109/ICIP.2005.1530609

[109] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations," *Color Research and Application,* vol. 30, pp. 21-30, 2005. doi: 10.1002/col.20070

[110] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 23, pp. 947-963, 2001.

[111] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, pp. 1075-1088, 2003.

[112] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-Tuned Salient Region Detection," *Computer Vision and Pattern Recognition, Proceedings of the IEEE Conf. on,* vol. 33, pp. 1536-1542, 2009. doi: 10.1109/CVPR.2009.5206596

[113] S. Ahn and J. Chong, "Segmenting a Noisy Low-Depth-of-Field Image Using Adaptive Second-Order Statistics," *IEEE Signal Processing Letters,* vol. 22, pp. 275-278, 2015. doi: 10.1109/LSP.2014.2357792

[114] R. Fielding, *Techniques of Special Effects of Cinematography*, 4 ed.: Focal Press, 1985. ISBN: 978-0240512341

[115] P. Favaro and S. Soatto, "A Geometric Approach to Shape from Defocus," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 406-417, 2005. doi: 10.1109/TPAMI.2005.43

[116] P. van Walree. (2002, June). *Depth of Field*, 2002. Available: http://toothwalker.org/optics/dof.html

[117] P. Grossmann, "Depth from Focus," *Pattern Recognition Letters,* vol. 5, pp. 63-69, 1987. doi: 10.1016/0167-8655(87)90026-2

[118] M. Subbarao, "Parallel Depth Recovery By Changing Camera Parameters," in *Computer Vision, Proceedings of the 2nd International Conference on*, 1988, pp. 149-155. doi: 10.1109/CCV.1988.589986

[119] M. Subbarao, "Determining Distance from Defocused Images of Simple Objects," State University of New-York, Stony Brook, NY 11794-2350, USA, 1989.

[120] B. K. P. Horn, *Robot Vision*: The MIT Press, 1986. ISBN: 978-0262081597

[121] P. Favaro and S. Soatto, *3-D Shape Estimation and Image Restoration: Exploiting Defocus and Motion-Blur*: Springer, 2006. ISBN: 1846281768

[122] D. Kundur and D. Hatzinakos, "Blind Image Deconvolution Revisited," *IEEE Signal Processing Magazine,* vol. 13, pp. 61-63, 1996. doi: 10.1109/79.543976

[123] G. R. Ayers and J. C. Dainty, "Iterative Blind Deconvolution Method and its Applications," *Optics Letters,* vol. 13, pp. 547-549, 1988. doi: 10.1364/OL.13.000547

[124] M. F. Fahmy, G. M. A. Raheem, U. S. Mohamed, and O. F. Fahmy, "A New Fast Iterative Blind Deconvolution Algorithm," *Journal of Signal and Information Processing,* vol. 3, pp. 98-108, 2012. doi: 10.4236/jsip.2012.31013

[125] A. Levin, R. Fergus, F. Durand, and W. T. Freemanm, "Image and Depth from a Conventional Camera with a Coded Aperture," in *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH*, 2007. doi: 10.1145/1276377.1276464

[126] M. F. Lindenbaum, M. and A. M. Bruckstein, "On Gabor's Contribution to Image Enhancement," *Pattern Recognition,* vol. 27, pp. 1-8, 1994. doi: 10.1016/0031-3203(94)90013-2

[127] Q. Wei, "Converting 2D to 3D: A Survey," TUDelft, Delft, Netherlands, 2005.

[128] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," in *Computer Vision, Proceeding of the 6th International Conference on*, Bombay, 1998, pp. 839-846. doi: 10.1109/ICCV.1998.710815

[129] K. N. Chaudhury, "Acceleration of the Shiftable O(1) Algorithm for Bilateral Filtering and Nonlocal Means," *Image Processing, IEEE Transactions on,* vol. 22, pp. 1291-1300, 2013. doi: 10.1109/TIP.2012.2222903

[130] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using Optimization," in *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH*, NY, USA, 2004, pp. 689-694. doi: 10.1145/1186562.1015780

[131] P. Heckbert, "Color Image Quantization for Frame Buffer Display," *ACM SIGGRAPH Computer Graphics,* vol. 16, pp. 297-307, 1982. doi: 10.1145/965145.801294

[132] P. Perona and J. Malik, "Scale-space and Edge Detection using Anisotropic Diffusion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 12, pp. 629-639, 1990. doi: 10.1109/34.56205

[133] E. S. L. Gastal and M. M. Oliveira, "Domain Transform for Edge-Aware Image and Video Processing," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2011,* vol. 30, 2011. doi: 10.1145/2010324.1964964

[134] K. N. Chaudhury, D. Sage, and M. Unser, "Fast O(1) Bilateral Filtering using Trigonometric Range Kernels," *Image Processing, IEEE Transactions on,* vol. 20, pp. 3376-3382, 2011. doi: 10.1109/TIP.2011.2159234

[135] F. Porikli, "Constant Time O(1) Bilateral Filtering," in *Computer Vision and Pattern Recognition (CVPR), Proceedings of the IEEE Conference on*, Anchorage, AK, 2008, pp. 1-8. doi: 10.1109/CVPR.2008.4587843

[136] C.-C. Cheng, C.-T. Li, and L.-G. Chen, "A Novel 2Dd-to-3D Conversion System Using Edge Information," *Consumer Electronics, IEEE Transactions on,* vol. 56, pp. 1739-1745, 2010. doi: 10.1109/TCE.2010.5606320

[137] C. Rother, "A New Approach for Vanishing Point Detection in Architectural Environments," in *British Machine Vision Conference (BMVC), Proceedings of the*, University of Bristol, UK, 2000, pp. 382-391. doi: 10.1.1.107.3980

[138] V. Cantoni, L. Lombardi, M. Porta, and N. Sicard, "Vanishing Point Detection: Representation Analysis and New Approaches," in *Image Analysis and Processing, Proceedings of 11th International Conference on*, Palermo, Italy (Insular), 2001, pp. 90-94. doi: 10.1109/ICIAP.2001.956990

[139] A. Almansa, A. Desolneux, and S. Vamech, "Vanishing Point Detection without Any A Priori Information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, pp. 502-507, 2003. doi: 10.1109/TPAMI.2003.1190575

[140] Y.-M. Tsai, Y.-L. Chang, and L.-G. Chen, "Block-based Vanishing Line and Vanishing Point Detection for 3D Scene Reconstruction," in *Intelligent Signal Processing and Communications (ISPACS), Proceedings of the International Symposium on*, Tottori, Japan, 2006, pp. 586-589.
doi: 10.1109/ISPACS.2006.364726

[141] J. Zhang and H.-H. Nagel, "Texture-Based Segmentation of Road Images," in *Intelligent Vehicles '94 Symposium, Proceedings of the*, 1994, pp. 260-265. doi: 10.1109/IVS.1994.639516

[142] J. D. Crisman and C. E. Thorpe, "UNSCARF - A Color Vision System for the Detection of Unstructured Roads," in *Robotics and Automation, Proceedings of IEEE International Conference on*, Sacramento, CA, 1991, pp. 2496-2501. doi: 10.1109/ROBOT.1991.132000

[143] C. Rasmussen, "Combining Laser Range, Color, and Texture Cues for Autonomous Road Following," in *Robotics and Automation, Proceedings of IEEE International Conference on*, 2002, pp. 4320-4325. doi: 10.1109/ROBOT.2002.1014439

[144] C. Coelho, M. Straforini, and M. Campani, "Using Geometrical Rules and a Priori Knowledge for the Understanding of Indoor Scenes," in *British Machine Vision Conference, Proceedings of the*, Oxford, 1990, pp. 229-234. doi: 10.5244/C.4.41

[145] Q. Wu, W. Zhang, T. Chen, and B. V. K. V. Kumar, "Prior-based Vanishing Point Estimation Through Global Perspective Structure Matching " in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Proceedings of the*, Dallas, TX, 2010, pp. 2110-2113. doi: 10.1109/ICASSP.2010.5495155

[146] P. Bellutta, G. Collini, A. Verri, and V. Torre, "3D Visual Information from Vanishing Points," in *Interpretation of 3D Scenes, Proceedings of the Workshop on*, Austin, TX, 1989, pp. 41-49. doi: 10.1109/TDSCEN.1989.68100

[147] J.-Y. Zhang, Y. Chen, and X.-X. Huang, "Edge Detection of Images Based on Improved Sobel Operator and Genetic Algorithms," in *International Conference on Image Analysis and Signal Processing (IASP) 2009, Proceedings of the*, Taizhou, 2009, pp. 31-35. doi: 10.1109/IASP.2009.5054605

[148] F. Y. Shih, "Image Processing and Mathematical Morphology: Fundamentals and Applications," 2 ed: CRC Press, 2009, pp. 184-192. ISBN: 978-1420089431

[149] P. V. C. Hough, "Method and Means for Recognizing Complex Patterns," United States of America Patent, 1962.

[150] S. R. Deans, "Hough Transform from the Radon Transform," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 3, pp. 185-188, 1981.

[151] A. Desolneux, L. Moisan, and J.-M. Morel, "Edge Detection by Helmholtz Principle," *Journal of Mathematical Imaging and Vision,* vol. 14, pp. 271-284, 2001. doi: 10.1023/A:1011290230196

[152] F. A. Andaló, G. Taubin, and S. Goldenstein, "Detecting Vanishing Points by Segment Clustering on the Projective Plane for Single-View Photogrammetry," presented at the Information Forensics and Security (WIFS), 2010 IEEE International Workshop on, Seattle, WA, 2010. doi: 10.1109/WIFS.2010.5711453

[153] R. O. Duda and P. E. Hart, "Use of the Hough Transform to Detect Lines and Curves in Pictures," *Comm. ACM,* vol. 15, pp. 11-15, 1972.

[154] J. A. Shufelt, "Performance Evaluation and Analysis of Vanishing Point Detection Techniques," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 21, pp. 282-288, 1999. doi: 10.1109/34.754631

[155] T. Tuytelaars, L. Van Gool, M. Proesmans, and T. Moons, "The Cascaded Hough Transform as an Aid in Aerial Image Interpretation," in *6th International Conference on Computer Vision (ICCV), Proceedings of the*, Bombay, India, 1998, pp. 67-72. doi: 10.1109/ICCV.1998.710702

[156] S. T. Barnard, "Interpreting Perspective Images," *Artificial Intelligence,* vol. 21, pp. 435-462, 1983. doi: 10.1016/S0004-3702(83)80021-6

[157] M. J. Magee and J. K. Aggarwal, "Determining Vanishing Points from Perspective Images," *Computer Vision, Graphics and Image Processing,* vol. 26, pp. 256-267, 1984. doi: 10.1016/0734-189X(84)90188-9

[158] L. Quan and R. Mohr, "Determining Perspective Structures using Hierarchical Hough Transform," *Pattern Recognition Letters,* vol. 9, pp. 279-286, 1989. doi: 10.1016/0167-8655(89)90006-8

[159] H. Nakatami, S. Kimura, and O. Saito, "Extraction of Vanishing Point and its Application to Scene Analysis Based on Image Sequence," in *Pattern Recognition, Proceeding of the 5th International Conference on*, 1980, pp. 370-372.

[160] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt. II," *Psychological Research,* vol. 4, pp. 301-350, 1923. doi: 10.1007/bf00410640

[161] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM,* vol. 24, pp. 381-395, 1981. doi: 10.1145/358669.358692

[162] A. Desolneux, L. Moisan, and J.-M. More, "A Grouping Principle and Four Applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 25, pp. 508-513, 2003. doi: 10.1109/TPAMI.2003.1190576

[163] T. S. Lee, "Image Representation using 2D Gabor Wavelets," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 18, pp. 959-971, 1996. doi: 10.1109/34.541406

[164] J. M. Coughlan and A. L. Yuille, "Manhattan World: Orientation and Outlier Detection by Bayesian Inference," *Neural Computation,* vol. 15, pp. 1063-1088, 2003. doi: 10.1162/089976603765202668

[165] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, *et al.*, "Depth Image-Based Rendering With Advanced Texture Synthesis for 3-D Video," *Multimedia, IEEE Transactions on,* vol. 13, pp. 453-465, 2011.

[166] C. Fehn, "Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV," in *Stereoscopic Displays and Virtual Reality Systems XI, SPIE Proccedings of*, San Jose, CA, 2004. doi: 10.1117/12.524762

[167] L.-M. Po, X. Xu, Y. Zhu, S. Zhang, K.-W. Cheung, and C.-W. Ting, "Automatic 2D-to-3D Video Conversion Technique based on Depth-from-Motion And Color Segmentation," in *IEEE 10th International Conference on Signal Processing (ICSP), Proceedings of the*, Beijing, 2010, pp. 1000-1003. doi: 10.1109/ICOSP.2010.5655850

[168] M. T. Pourazad, P. Nasiopoulos, and R. K. Ward, "Generating the Depth Map from the Motion Information of H.264-Encoded 2D Video Sequence," *EURASIP Journal on Image and Video Processing,* vol. 2010, 2010. doi: 10.1155/2010/108584

[169] A. J. Woods, T. Docherty, and R. Koch, "Image Distortions in Stereoscopic Video Systems," in *Stereoscopic Displays and Applications IV, Proceedings of SPIE*, San Jose, CA, 1993. doi: 10.1117/12.157041

[170] H.-K. Hong, M. S. Ko, Y.-H. Seo, D.-W. Kim, and J. Yoo, "3D Conversion of 2D Video Encoded by H.264," *Journal of Electrical Engineering & Technology (JEET),* vol. 7, pp. 990-1000, 2012. doi: 10.5370/JEET.2012.7.6.990

[171] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*. Oxford, United Kingdom: Oxford Scholarship Online, 2008.

[172] W. J. Tam, G. Alain, L. Zhang, T. Martin, R. Renaud, and D. Wang, "Smoothing Depth Maps for Improved Steroscopic Image Quality " in *Three-Dimensional TV, Video, and Display III, Proceedings of SPIE* vol. 5599, B. Javidi and F. Okano, Eds., ed. Philadelphia, U.S.A., 2004, pp. 162-172.

[173] K. Moustakas, D. Tzovaras, and M. G. Strintzis, "Stereoscopic Video Generation Based on Efficient Layered Structure and Motion Estimation From a Monoscopic Image Sequence," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 15, pp. 1065-1073, 2005. doi: 10.1109/TCSVT.2005.852401

[174] W.-Y. Chen, Y.-L. Chang, S.-F. Lin, and L.-F. Ding, "Efficient Depth Image Based Rendering with Edge Dependent Depth Filter and Interpolation," in *IEEE International Conference on Multimedia and Expo (ICME), Proceedings of the*, 2005, pp. 1314-1317. doi: 10.1109/ICME.2005.1521671

[175] C. Guillemot and O. Le Meur, "Image Inpainting: Overview and Recent Advances," *IEEE Signal Processing Magazine,* vol. 31, pp. 127-144, 2013. doi: 10.1109/MSP.2013.2273004

[176] M. Bertalmio, C. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *SIGGRAPH 2000, Proceedings of*, 2000, pp. 417–424.

[177] A. Criminisi, P. Pérez, and K. Toyama, "Region Filling and Object Removal by Exemplar-Based Image Inpainting," *IEEE Transactions on Image Processing,* vol. 13, pp. 1200-1212, 2004. doi: 10.1109/TIP.2004.833105

[178] A. A. Efros and T. K. Leung, "Texture Synthesis by Non-parametric Sampling," in *Computer Vision, 1999, The Proceedings of the Seventh IEEE International Conference on*, Corfu, Greece, 1999. doi: 10.1109/ICCV.1999.790383

[179] G. Nur, V. De Silva, H. Kodikara Arachchi, A. F. Kondoz, W., M. O. Martinez-Rach, and S. Dogan, "Sensitivity of the HVS for Binocular disparity Cue in 3D Displays under Different Ambient Illumination Conditions," in *International Conference on Consumer Electronics (ICCE), Proceedings of the 2012 IEEE*, 2012, pp. 459-460. doi: 10.1109/ICCE.2012.6161943

[180] S. Reeve and J. Flock. (2010). *Basic Principles of Stereoscopic 3D*: British Sky Broadcasting Group, 2010. Available: http://www.sky.com/shop/PDF/3D/Basic_Principles_of_Stereoscopic_3D_v1.pdf

[181] J. B. Kaiser, *Make Your Own Stereo Pictures*. New York: The Macmillan Company, 1955. ISBN: B0000CJ9P0

[182] W. Rollman, "Zwei Neue Stereoskopische Methoden," *Annalen der Physik,* vol. 166, pp. 186-187, 1853. doi: 10.1002/andp.18531660914

[183] E. Dubois, "A Projection Method To Generate Anaglyph Stereo Images," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Proceedings of the 2001 IEEE*, Salt Lake City, UT, 2001, pp. 1661-1664. doi: 10.1109/ICASSP.2001.941256

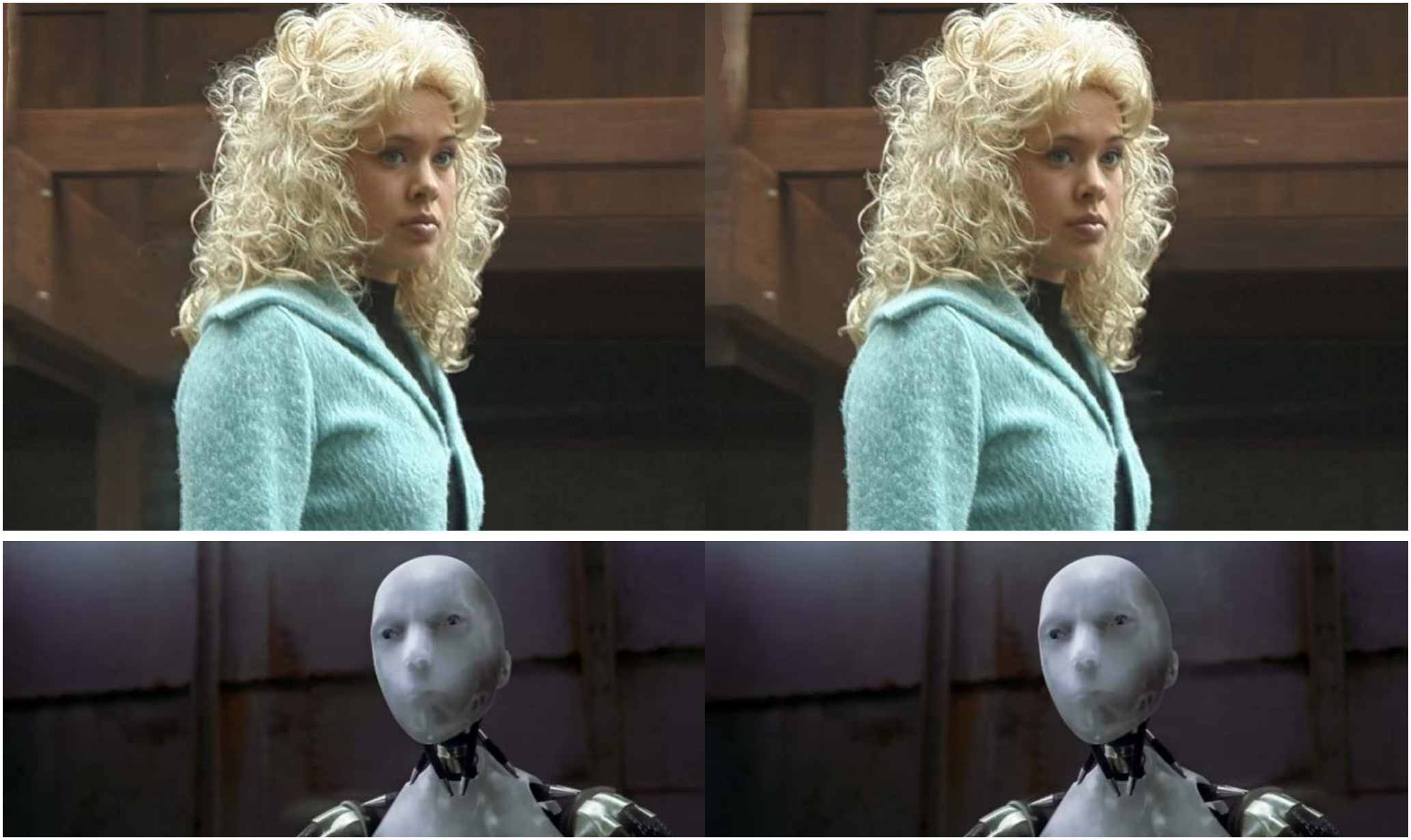
[184] D. F. McAllister and Z. Zhang, "A Uniform Metric for Anaglyph Calculation," in *Stereoscopic Displays and Virtual Reality Systems XIII, Proceedings of*, San Jose, CA, 2006. doi: 10.1117/12.642898

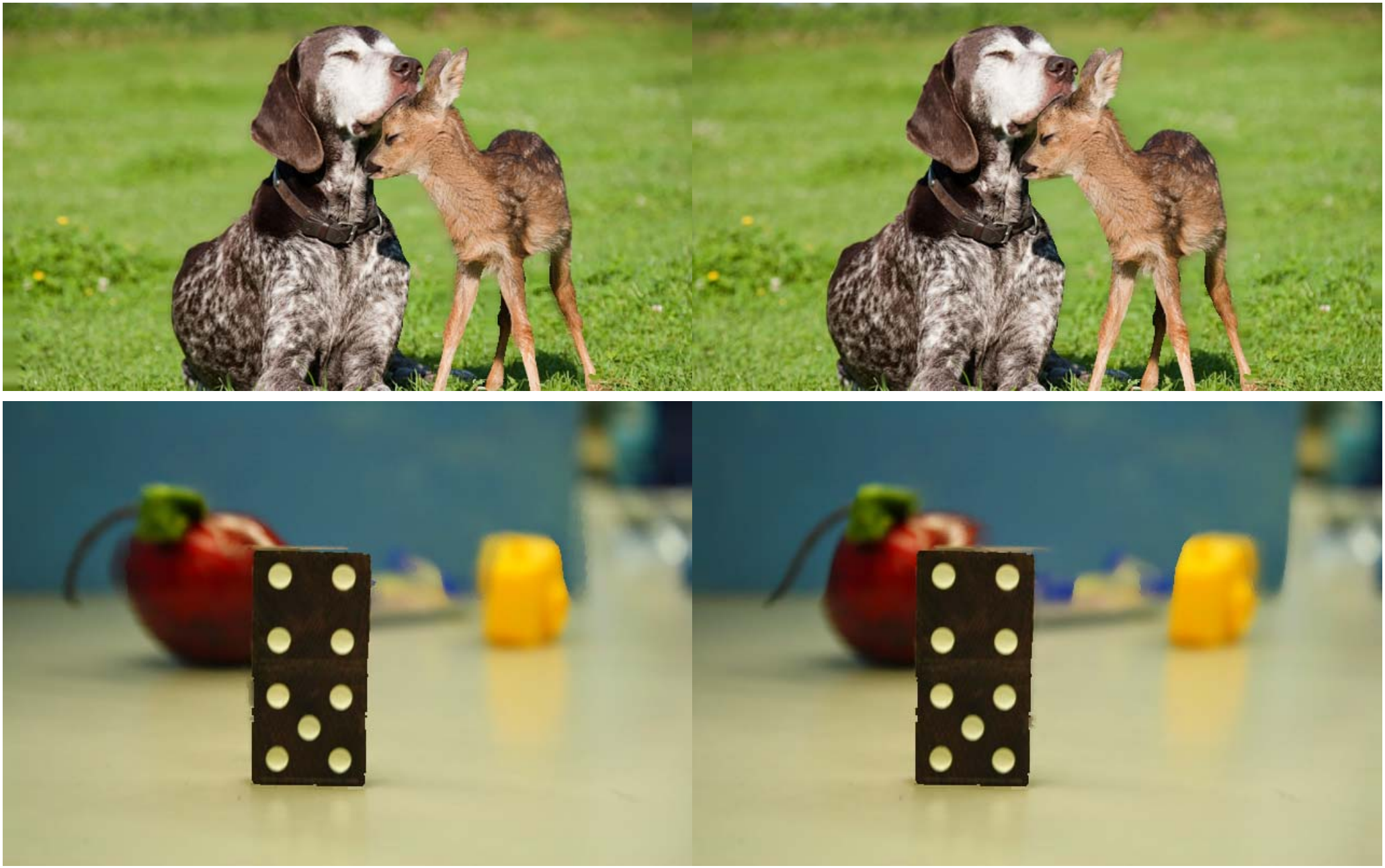
[185] D. F. McAllister, Y. Zhou, and S. Sullivan, "Methods for Computing Color Anaglyphs," in *Stereoscopic Displays and Applications XXI, Proceedings of the SPIE*, San Jose, California, 2010. doi: 10.1117/12.837163

X. APPENDIX


Jake Olson Studios Copyright 2013


Jake Olson Studios Copyright 2013






Coca-Cola