Multi–modal Scene Reconstruction using Perceptual Grouping Constraints

Nicolas Pugeault
University of Edinburgh
npugeaul@inf.ed.ac.uk

Florentin Wörgötter
University Göttingen
worgott@chaos.gwdg.de

Norbert Krüger
Aalborg University Copenhagen
nk@imi.aau.dk

Abstract

In this work we propose a scheme integrating perceptual grouping into stereopsis to reduce the ambiguity of those early processes. We propose a simple perceptual grouping algorithm that – in addition to the geometric information – makes use of a novel multi–modal affinity measure between local primitives. We then use this group information to 1) disambiguate the stereopsis by enforcing that stereo matches preserve groups; and 2) correct the reconstruction error due to the image pixel sampling using a linear interpolation over the groups. We show quantitative and qualitative demonstrations of those processes on a variety of sequences.

1. Introduction

We propose in this paper an approach using feedback between two mid–level processes, namely perceptual grouping and stereopsis to reduce the ambiguity omnipresent at this level of processing. We base our framework on a novel image representation based on multi–modal local image descriptors called *primitives*, introduced by [21] and applied to stereo by [20]. In this work, we will focus on primitives describing line structures, and we propose a perceptual grouping mechanism which makes use of this rich multi–modal information.

Perceptual grouping can be divided in two tasks: 1) defining an affinity measure between primitives and use it to build a graph of the connectedness between the primitives, and 2) extracting groups, which are the connected components of this graph. We will only define the affinity measure between primitives, and not extract the groups themselves explicitly, as we only need the local grouping information for a primitive to apply the correction mechanisms we propose in this paper. Similar affinity measures have been proposed by [27, 26], which formalised a good continuation constraint, or [9] which included the intensity on each side of the curve into a Bayesian formulation of grouping. Yet in this paper we propose a multi–modal similarity measure, composed of phase, colour and optical flow measurement, and combine it with a classical good continuation criterion forming a novel multi–modal definition of the affinity between primitives. Note that an explicit description of the groups could be extracted easily using a variety of techniques including: normalised [34] or average cuts [32], affinity normalisation [27], dynamic programming [33], *etc*.

The interest of using perceptual organisation in the spatial and temporal domains has been outlined by [31]. Here, we will study how this perceptual grouping information can be used to disambiguate stereopsis and 3D reconstruction using primitives. If we assume that a contour of the image is likely to be a projection of a contour of the 3D scene, then we can expect each 3D contour of the scene to project as a 2D contour on each camera plane (except in the case of occlusion). Conversely, this also implies that any contour in one image has a corresponding contour in the second image (or it is occluded). Thus we will propose an *external* stereo confidence which estimates how well primitives that are part of the same group agree with a putative stereo–match. This allows to discard a large number of potential stereo–correspondences hence reducing the ambiguity of the stereo matching and of the scene reconstruction processes.

We will test this scheme with four different calibrated stereo sequences, illustrated in figure 1. For sequences (a) (b) and (c) we have depth values obtained from a range scanner. Ten different frames from those three sequences were used for quantification in this paper. Sequence (d) was recorded outdoors in a moving car. for which we will show qualitative results.

The novel contributions of this paper are

- a 2D grouping that uses geometric and appearance based information,
- using the 2D grouping for improving stereo matching from a very local level (in contrast to, e.g., [30], where more elaborate features, like ribbons, were considered),
- applying an interpolation method that leads to more reliable estimates of 3D position and 3D-orientation.

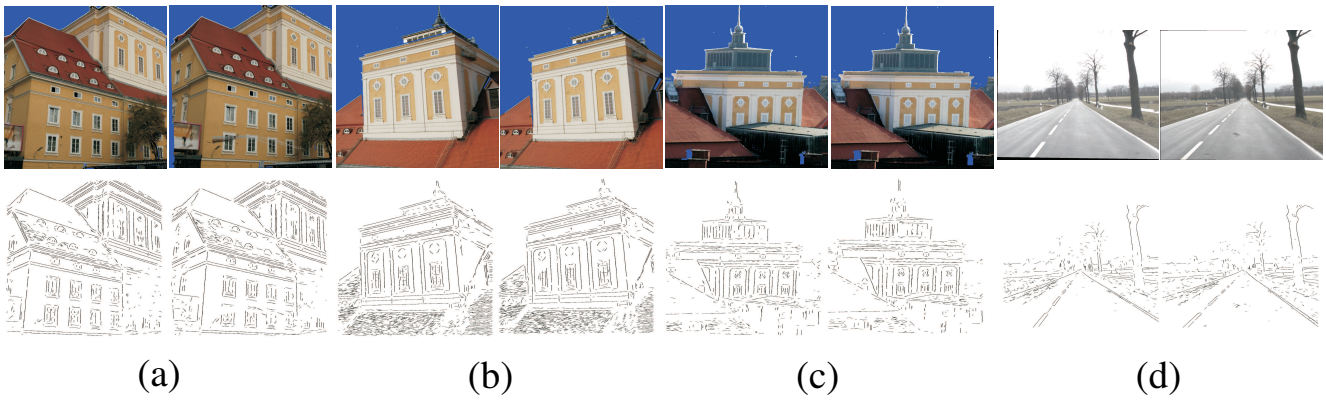


Figure 1. The four sequences on which we tested our approach.

The grouping is part of an early cognitive vision framework including ego-motion estimation and temporal accumulation (for an outline see [37]).

The paper is structured as follows: Section 2 will present the image primitives on which we are basing our processing. In section 3, we define the affinity between two primitives. In section 4 we present a stereo–matching process based on primitives similar to [20]. Then in section 5 we propose a simple scheme to 1) increase the reliability of matching and 2) smooth the reconstruction of a stereo sequence using information gained from the perceptual grouping defined earlier.

2. 2D–primitives

Numerous feature detectors exist in the literature (see [22] for a review). Each feature based approach can be divided into an interest point detector (e.g. [3, 4]) and a descriptor describing a local patch of the image at this location, that can be based on histograms (e.g. [6, 22]), spatial frequency [28], local derivatives [15, 13, 1] steerable filters [36], or invariant moments ([23]). In [22] these different descriptors have been compared, showing a best performance for SIFT-like descriptors.

The primitives we will be using in this work are local, multi–modal edge descriptors that were introduced in [21]. In contrast to the above mentioned features these primitives focus on giving a semantically and geometrically meaningful description of the local image patch. The importance of such a semantic grounding of features for a general purpose vision front–end, and the relevance of edge–like structures for this purposes were discussed in [7].

The edge map and the local phase are computed using the monogenic signal (see [11]), although some other kind of filtering could alternatively be used (e.g., steerable filters [36]). The primitives are extracted sparsely at locations in the image that are the most likely to contain edges. This likelihood is computed using the intrinsic dimensionality measure proposed in [19]. The sparseness is assured using a classical winner take all operation, insuring that the generative patches of the primitives do not overlap. Each of the primitive encodes the image information contained by a local image patch of a same size ρ as the kernel used by the filtering operation. Multi–modal information is gathered from this image patch, including the position $\boldsymbol{m}$ of the centre of the patch, the orientation θ of the edge, the phase ω of the signal at this point, the colour $\boldsymbol{c}$ sampled over the image patch on both sides of the edge and the local optical flow $\boldsymbol{f}$, computed using the classical Nagel algorithm (see [25]). Consequently a local image patch is described by the following multi–modal vector:

$$\boldsymbol{\pi} = (\boldsymbol{m}, \theta, \omega, \boldsymbol{c}, \boldsymbol{f}, \rho)^T \tag{1}$$

that we will name *primitive* in the following. The set of primitives describing the stereo images is called *image representation* and written $\boldsymbol{\mathcal{I}}^l$ and $\boldsymbol{\mathcal{I}}^r$ for the images from the left and right camera. The image representation extracted from one image is illustrated in figure 2.

Note that these primitives are of lower dimensionality than, e.g., SIFT (10 vs. 128) and therefore suffer of a lesser distinctiveness. Nonetheless, we will show in section 4 that they are distinctive enough for a reliable stereo matching if the epipolar geometry of the cameras is known. Advantageously, the rich information carried by the 2D–primitives can be reconstructed in 3D, providing a more complete scene representation. Having geometrical meaning for the primitive allows to describe the relation between proximate primitives in terms of perceptual grouping.

3. Perceptual Grouping of 2D–Primitives

Decades ago, the Gestalt psychologists proposed a series of axioms describing the way the human visual system binds together features in an image (see [16, 35, 17]). This process is generally called *perceptual grouping* the Gestalt psychologists proposed that it was driven properties like proximity, good continuation, similarity, symmetry, amongst others. More recently, psychophysical experiments measured the impact of different cues for percep-

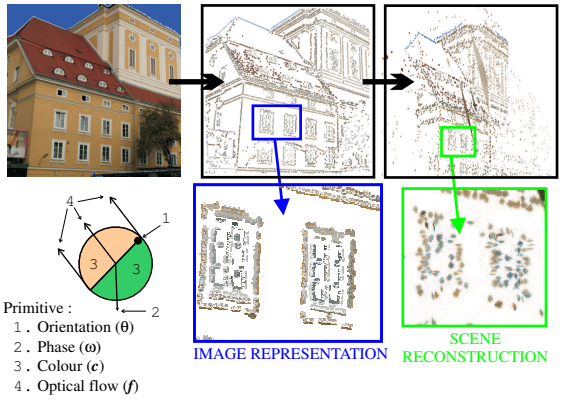


Figure 2. Illustration of the primitive extraction process from a video sequence. The figure shows one image from the sequence (a) from figure 1, on the right, then the 2D–primitives extracted from this image (see section 2), and finally the 3D–primitives reconstructed from the stereo–matches as described in section 4. The bottom row shows a description of the graphic representation of the 2D–primitives, as well as a magnification of the image representation and the reconstructed entities. Note that the structure reconstructed is quite far from the cameras, leading to a certain imprecision in the reconstruction of the 3D–primitives. We will propose a simple scheme addressing this problem in section 5.3

tual grouping (see, e.g., [12]). Furthermore, Brunswik and Kamiya [2] proposed that those processes should be related to statistics of natural images, which has been recently confirmed by several studies [18, 8, 14].

We previously defined the primitives as local edge descriptors, and that a group of primitives describe a contour of the image. The Gestalt rule of *proximity* implies that primitives that are closer to one another are most likely to lie on the same contour. According to the Gestalt rule of *good continuation*, we will consider that contours in the image are smooth, and therefore that two proximate primitives in a group will be nearly either collinear or co–circular. In this formulation, a strong inflexion in a contour will lead this contour to be described as *two* groups joining at the inflection point. Furthermore the position and orientation of primitives that are part of a group are the local tangents to the contour described by this group. Finally, the rule of *similarity* states that primitives that are similar (in terms of the colour, phase and optical flow modalities) are most likely to be grouped together. Also, we would expect such properties as colour on both side of a contour to change smoothly along this contour.

The two first cues are joined into a *Geometric constraint* that we describe in section 3.1 and the multi–modal similarity cue is detailed in section 3.2. These two measures are combined into an overall affinity measure that we describe in section 3.3.

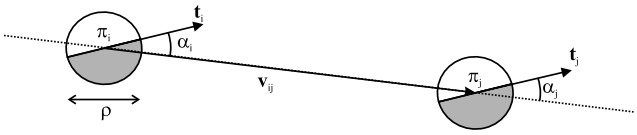


Figure 3. Illustration of the values used for the collinearity computation. If we consider two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$, then the vector between the centres of these two primitives is written $\boldsymbol{v}_{ij}$, and the orientations of the two primitives are designated by the vectors t_i and t_j, respectively. The angle formed by $\boldsymbol{v}_{ij}$ and t_i is written α_i, and between $\boldsymbol{v}_{ij}$ and t_j is written α_j. ρ is the radius of the image patch used to generate the primitive.

3.1. Geometric constraint

If we consider two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ in $\mathcal{I}$, then the likelihood that they both describe the same contour can be formulated as a combination of three basic constraints on their relative position and orientation — see figure 3.

Proximity (c_p []):

$$c_p\left[g_{i,j}\right] = 1 - e^{-\max\left(1-\frac{||\mathbf{v_{i,j}}||}{\rho\tau},0\right)} \tag{2}$$

Here, ρ stands for the radius of the the primitives in pixels. $\rho\tau$ is the size of the neighbourhood considered in pixels. $\|\mathbf{v_{i,j}}\|$ is the distance in pixels separating the centres of the two primitives.

Collinearity (c_{co} []):

$$c_{co}\left[g_{i,j}\right] = 1 - \left|\sin\left(\frac{|\alpha_i| + |\alpha_j|}{2}\right)\right| \tag{3}$$

Here α_i and α_j are the angles between the line joining the two primitives centres and the orientation of, respectively, $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$.

Co–circularity (c_{ci} []):

$$c_{ci}\left[g_{i,j}\right] = 1 - \left|\sin\left(\frac{\alpha_i + \alpha_j}{2}\right)\right| \tag{4}$$

The combination of those three criteria forms the *geometric* affinity measure:

$$\mathbf{G}_{i,j} = \sqrt[3]{c_e\left[g_{i,j}\right] \cdot c_{co}\left[g_{i,j}\right] \cdot c_{ci}\left[g_{i,j}\right]} \tag{5}$$

where $\mathbf{G}_{i,j}$ is the geometric affinity between two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$. This affinity represent the likelihood for a curve having for tangents those two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_i$ to be an actual contour of the scene.

3.2. Multi–modal Constraint

Effectively, the more similar are the modalities between two primitives, the more likely are those two primitives to lie on the same contour. Note that [8] already proposed to use the intensity as a cue for perceptual grouping, yet here

we use a combination of the phase, colour and optical flow modalities of the primitives to decide if they describe the same contour:

$$\mathbf{M}_{i,j} = 1 - w_\omega d_\omega(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) - w_{\boldsymbol{c}} d_{\boldsymbol{c}}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) - w_{\boldsymbol{f}} d_{\boldsymbol{f}}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) \quad (6)$$

where d_ω is the phase distance, $c_{\boldsymbol{c}}$ the colour distance and $c_{\boldsymbol{f}}$ the optical flow distance between the two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$. These metrics are similar to the ones used in [29, 20]. w_ω, $w_{\boldsymbol{c}}$ and $w_{\boldsymbol{f}}$ are the relative weight of the modalities, such that $w_\omega + w_{\boldsymbol{c}} + w_{\boldsymbol{f}} = 1$.

3.3. Primitive Affinity

The overall affinity between all primitives in an image is formalised as a matrix $\boldsymbol{A}$, where $\boldsymbol{A}_{i,j}$ holds the affinity between the primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$. We define this affinity from equations (5) and (6), such that 1) two primitives complying poorly with the good continuation rule have an affinity close to zero; and 2) two primitives complying with the good continuation rule yet strongly dissimilar will have only an average affinity. The affinity is formalised as follows:

$$c[g_{i,j}] = \boldsymbol{A}_{i,j} = \sqrt{\mathbf{G}\left(\alpha \mathbf{G}_{i,j} + (1-\alpha)\mathbf{M}_{i,j}\right)} \quad (7)$$

where α is the weighting of geometric and multi–modal (*i.e.* phase, colour and optical flow) information in the affinity. A setting of $\alpha = 1$ implies that only geometric information (proximity, collinearity and co-circularity) is used, while $\alpha = 0$ indicates that geometric and multi–modal information are evenly mixed. The groups generated for the left and right frames for each sequence are drawn in figure 1, bottom row. Dark lines describe strings of grouped primitives. One can see in those images that the major contours of the images are adequately described.

4. Stereopsis using 2D–primitives

Classical stereopsis allows reconstructing a 3D point from two corresponding stereo points. A review of stereo–algorithms was presented in [24], dense two frames stereo algorithms were also compared in [5]. In these papers the different algorithms were compared on mainly artificial images, with a disparity d that ranges in $0 \leq d \leq 16$. In this work we make use of a sparse, feature based representation, applied on high resolution video sequences of natural scenes, where the ground truth was obtained using a range scanner. The allowed disparity range for these scenes is $0 \leq d \leq 200$, leading to a comparable level of ambiguity (*i.e.* between 10 and 20 candidates depending on the primitive being matched).

The stereopsis used for this paper is a simple local winner–take–all scheme: all primitives in the right image that lie on the epipolar line are *potential correspondences* and their individual likelihood is set as their multi–modal similarity with the original primitive in the left image. Then the most similar primitive is taken as the most likely correspondence. The multi–modal distance between two primitives is defined as a linear combination of the modal distances between the two primitives:

$$d_m(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = \sum_m w_m d_m(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) \quad (8)$$

where w_m is the relative weighting of the modality m, with $\sum_m w_m = 1$ (we use distance functions for the modalities that are similar to the ones proposed in [29, 20]).

In figure 6(a) the ROC curves showing the performance of the stereo–matching when using as likelihood estimation the similarities in each of the modalities held by a primitive, alongside with the performance of the multi–modal distance proposed in equation (8). We can see that: 1) all modalities offer a discrimination better than chance between correct and erroneous correspondences; and 2) the multi–modal distance offers a better discrimination than the individual modalities. In this figure we can see that the colour modality is a particularly strong discriminant for stereopsis. This is explained by the fact that the hue and saturation are sampled on each side of the edge, leading to a 4–dimensional modality, where phase and orientation are only 1–dimensional and optical flow is 2–dimensional (albeit the aperture problem reduces it to one effective dimension: the normal flow). On the other hand the poor performance of the optic flow modality could be explained by the relative simplicity of the motion in this scene: a pure forward translation of the camera, with no moving object. Therefore, we would expect the performance of individual modalities to vary depending on the scenario, and the robustness of the multi–modal constraint could be further enhanced by a contextual weighting. Nevertheless, in a variety of scenarios the use of a static weighting proved robust enough to obtain reliable stereopsis.

Moreover, by making use of the rich semantic information carried by the primitives, the stereopsis yield a set of geometrically meaningful entities rather than an mere disparity map We call the reconstructed entities 3D–primitives $\boldsymbol{\Pi}$:

$$\boldsymbol{\Pi} = (\boldsymbol{M}, \boldsymbol{\Theta}, \Omega, \boldsymbol{C})^T \quad (9)$$

where $\boldsymbol{M}$ is the location in space, $\boldsymbol{\Theta}$ is the 3D orientation of the edge, Ω is the phase across this edge, and $\boldsymbol{C}$ holds the colour information for this edge — see attached material. In figure 7(a) we show the 3D–primitives that were reconstructed after a stereo–matching based on the multi–modal confidence from equation (8).

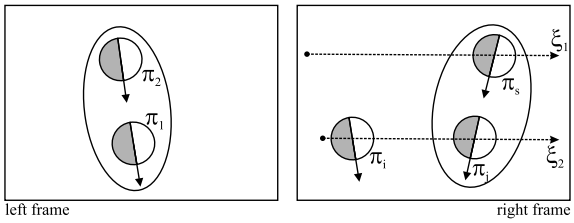


Figure 4. The BSCE criterion: Let $\boldsymbol{\pi}_1$ be a primitive in the left frame forming a group with a second primitive $\boldsymbol{\pi}_2$. $\boldsymbol{\pi}_2$ has a stereo correspondence $\boldsymbol{\pi}_s$ in the right image. Both $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ in the right image lie on the epipolar line ξ_1 of $\boldsymbol{\pi}_1$; hence these two primitives are both putative correspondences of $\boldsymbol{\pi}_1$. Furthermore, the primitive $\boldsymbol{\pi}_i$ is clearly the most similar to $\boldsymbol{\pi}_1$ (due to a closer orientation), hence this stereo–correspondence $s_{1\to i}$ yield a higher multi–modal confidence than would, e.g. $s_{1\to j}$. Yet, when considering the BSCE criterion we realise that only the putative correspondence $\boldsymbol{\pi}_j$ forms a group $g_{j,s}$ with $\boldsymbol{\pi}_s$, conserving the group relation $g_{1,2}$ between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$.

5. Perceptual Grouping Constraints to Improve Stereopsis

In addition to their richness, primitives are very redundant along contours, and this redundancy allows us to use perceptual grouping to derive the following two constraints for the matching process:

Isolated primitives are likely to be unreliable: As primitives are extracted redundantly along the contours, conversely an isolated primitive is likely to be an artifact. Hence isolated primitives can be neglected.

Stereo consistency over groups: If a set of primitives forms a contour in the first image, the *correct correspondences* of these primitives in the second image also form a contour.

5.1. Basic Stereo Consistency Event (BSCE)

As explained in section 3, 2D–primitives represent local estimators of image contours. A constellation of those 2D–primitives describe the contour as a whole. Those contours are consistent over stereo, with the notable exception of partially occluded contours — see figure 1, bottom row. Hence, if two primitives describe a contour in one image then their correspondences in the second image should also describe the same contour, and those two 2D contours are the projection of the same 3D contour onto the two different optical planes. In section 3, we defined the likelihood for two primitives to describe the same contour as the affinity between these two primitives, hence we can rewrite the previous statement as:

Given two primitives $\boldsymbol{\pi}_i^l$ and $\boldsymbol{\pi}_j^l$ in $\boldsymbol{\mathcal{I}}^l$ and their respective correspondences $\boldsymbol{\pi}_n^r$ and $\boldsymbol{\pi}_p^r$ in a second image $\boldsymbol{\mathcal{I}}^r$; if $\boldsymbol{\pi}_i^l$ and $\boldsymbol{\pi}_j^l$ belongs to the same group in $\boldsymbol{\mathcal{I}}^l$ then $\boldsymbol{\pi}_n^r$ and $\boldsymbol{\pi}_p^r$ should also be part of a group in $\boldsymbol{\mathcal{I}}^r$. — see figure 4.

We call the conservation of the link between a pair of primitives in the stereo–correspondences of those primitives the *Basic Stereo Consistency Event* (BSCE).

This condition can then be used to test the validity of a stereo–hypothesis. Consider a primitive $\boldsymbol{\pi}_i^l$, and a stereo hypothesis:

$$s_{i\to n} : \boldsymbol{\pi}_i^l \to \boldsymbol{\pi}_n^r \tag{10}$$

and consider a neighbour $\boldsymbol{\pi}_j^l \in N(\boldsymbol{\pi}_i^l)$ of $\boldsymbol{\pi}_i^l$ such that the two primitives share an affinity $c\,[g_{i,j}]$. For this second primitive a stereo–correspondence $\boldsymbol{\pi}_p^n$ with a confidence of $c\,[s_{j\to p}]$ exists. We can then estimate how well the stereo–hypothesis $s_{i\to n}$ preserves the BSCE:

$$E(g_{i,j}, s_{i\to n}) = \begin{cases} \sqrt{c\,[s_{j\to p}] \cdot c\,[g_{i,j}]} & if\ c\,[g_{n,p}] > \varepsilon \\ -\sqrt{c\,[s_{j\to p}] \cdot c\,[g_{i,j}]} & else \end{cases} \tag{11}$$

In other words, considering a stereo–pair of primitives: the BSCE of a primitive in the first image with one of its neighbour is high if they share a strong affinity and if this second primitive creates a stereo–hypothesis such that the correspondences in the second image of both primitives *also* share a strong affinity. It is low if the stereo–correspondence of this primitive and the stereo–correspondences of other primitives part of the same group, do not form a group in the other image. This naturally extends the concept of group as defined in section 3 into the stereo domain.

5.2. Neighbourhood Consistency Confidence

Building on the formula (11), we can define how *the whole neighbourhood* of a primitive is consistent with a given stereo hypothesis.

The previous formula tells us how a 2D–primitive stereo correspondence is consistent with our knowledge of the set of stereo hypotheses for a second 2D–primitive, in its neighbourhood. Now, if we consider a primitive $\boldsymbol{\pi}_i^l$ and an associated stereo–correspondence $s_{i\to n}$, we can integrate this BSCE confidence over the neighbourhood of the primitive $\boldsymbol{\mathcal{N}}_i^l$ — as defined in section 3.3.

$$c_{ext}[s_{i\to n}] = \frac{1}{\#\boldsymbol{\mathcal{N}}_i^l} \sum_{\boldsymbol{\pi}_k^l \in \boldsymbol{\mathcal{N}}_i^l} E(\boldsymbol{\pi}_1^l, \boldsymbol{\pi}_k^l, s_{i\to n}) \tag{12}$$

Where $\#\boldsymbol{\mathcal{N}}_i^l$ is the size of the neighbourhood — *i.e.* the number of neighbours of $\boldsymbol{\pi}_1^l$ considered. We call this new confidence the *external confidence* in $s_{i\to n}$, as opposed to the internal confidence given by the multi–modal similarity between the 2D–primitives— equation (8). In figure 5, one can see that the correct correspondences have mostly positive external confidences, while incorrect ones have mainly negative values. Therefore, applying a threshold on the external confidence will remove stereo hypotheses that are inconsistent with their neighbourhood, and thus reduce the ambiguity of the stereo–matching. Note that selecting a

threshold higher than zero implies the removal of all the isolated primitives (as an isolated primitive has an external confidence of zero by definition).

Figure 6(b) shows ROC curves of the performance for varying thresholds on the multi–modal similarity. Each of the curve drawn shows the performance for different thresholds (respectively threshold values of $-0.6, -0.3, 0, +0.3$, and without threshold) applied to the external confidence prior to the ROC analysis. We can see from those results that applying a bias on the decision based on the external confidence is improving significantly the accuracy of the decision process. Depending on the type of selection process desired — very selective and reliable, or more lax, but yielding a denser set of correspondences — another threshold can be chosen. The best overall improvement seems to be reached for a threshold of -0.3 over the external confidence. Nonetheless, when we consider a case where very high reliability is required, a threshold of 0 (meaning discarding all primitives which are part of no group) might be preferred. Note that when a threshold is applied to the external confidence prior to the ROC analysis, the resulting curve do not reach the $(1, 1)$ point of the graph. This is normal as the threshold already remove some stereo–hypotheses even before the multi–modal confidence is considered.

The 3D–primitives reconstructed after such a scheme are shown in figure 7(b).

5.3. Interpolation in Space

One issue when reconstructing 3D structures from stereopsis is that the accuracy of the reconstructed entities is decreasing with the distance to the cameras, due to the pixel sampling of the images — see [10]. Figure 7(b) shows the reconstruction of the tree (along with the road markings) in sequence (d) — see figure 1. There we can see that, although all primitives describe the contour of the tree from the same point of view, their exact position and orientation in space vary, and they certainly do not form a contour in space.

Yet, we do know that the 2D–primitives they are reconstructed from a group in both stereo images (*c.f.* section 5 and figure 1 bottom row), and as such that they form a smooth continuous contour. Hence we can assume that they are the projection on the image planes of a smooth and continuous contour of the scene (except in some extreme cases and under rare viewpoints), and as such that the reconstructed 3D–primitives should also describe such a curve.

A common way of reducing such noise in the sampling of a smooth function is to use linear smoothing, hence we propose to apply it to the 3D–primitives. For each iteration n of this smoothing, the position $\boldsymbol{M}$ and orientation $\boldsymbol{\Theta}$ of the primitive $\boldsymbol{\Pi}_i^{(n)}$ are changed to the average between their previous values $\boldsymbol{\Pi}_i^{(n-1)}$ and values interpolated from the primitives reconstructed out of the two closest neighbours of the 2D–primitive in the images $I(\boldsymbol{\Pi}_j^{(n-1)}, \boldsymbol{\Pi}_k^{(n-1)})$.

$$\boldsymbol{M}_i^{(n)} = \frac{1}{2}\left(\boldsymbol{M}_i^{(n-1)} + I(\boldsymbol{M}_j^{(n-1)}, \boldsymbol{M}_k^{(n-1)})\right) \quad (13)$$

$$\boldsymbol{\Theta}_i^{(n)} = \frac{1}{2}\left(\boldsymbol{\Theta}_i^{(n-1)} + I(\boldsymbol{\Theta}_j^{(n-1)}, \boldsymbol{\Theta}_k^{(n-1)})\right) \quad (14)$$

Figure 7 illustrate the reconstructed 3D–primitives from the sequence (d) (*c.f.* figure 1). Note that it is necessary to choose a point of view sufficiently different from the one of the camera to highlight the reconstruction errors, while being sufficiently similar for the shapes of the scene to be recognisable. We chose a point of view located high on the right side of the scene, looking downwards at the road.

When comparing figures 7(a) and 7(b) we can see that a large number of outliers are discarded from the reconstructed 3D–primitives, leading to a cleaner description of the scene. Figure 7(c) shows the same part of the scene (d) after 3 iterations of the linear smoothing. The 3D–primitives forming the contour of the tree and the road markings are now smoothly aligned.

6. Conclusion

In this paper we defined an affinity relation between image primitives making use of the rich multi–modal information available. Therefore the resulting affinity measure encompass more than just the good continuation cue but also continuity in phase, colour and optical flow. We have illustrated that, on varied sequence, the resulting groups follow adequately the contours of the image. In a second part we proposed a simple measure of the conservation of those groups, and hence of the neighbourhood structure of a primitive, across stereo. Using this conservation we could formalise a contextual estimation of the likelihood of a stereo correspondence. We show that using this new external confidence measure in conjunction with a similarity measure we can improve significantly the performance of the stereo–matching process. Furthermore, we show that interpolation can be used over a group to correct the smoothness of the reconstructed representation.

Acknowledgement: We thank the company Riegl for the images with known ground truth used for sequence (a), (b) and (c). This work described in this paper was part of the European project ECOVISION.

References

[1] A. Baumberg. Reliable Feature Matching across Widely Separated Views. In *Proc. Conf. Computer Vision and Patter Recognition*, pages 774–781, 2000. 2

[2] E. Brunswick and J. Kamiya. Ecological cue validity of 'proximity' and other gestalt factors. *Journal of Psychology*, 66:20–32, 1953. 3

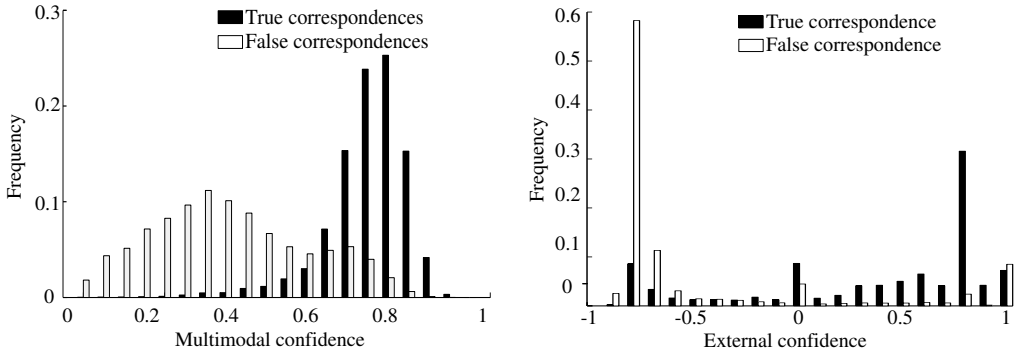
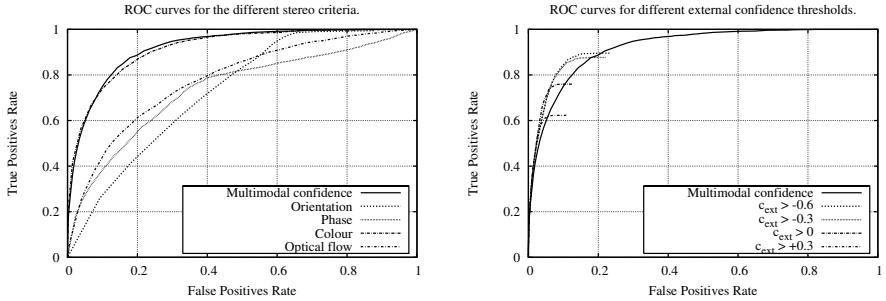


Figure 5. Distribution of multi–modal similarity and external confidence for correct (black bars) and false (white bars) correspondences. These data have been collected over 10 frames of the sequences (a), (b) and (c) — see figure 1.



(a) multi–modal similarity only

(b) multi–modal + external confidence

Figure 6. ROC curves for the performance of the multi–modal confidence to discriminate correct from erroneous correspondences. (a) Comparisons of the different modalities for stereo–matching (see for a discussion of the role of colour in the text). (b) Each curve stands for a the application of a different threshold over the external confidence, prior to the ROC analysis. Those curves represent the statistics over 10 frames of the two sequences with ground truth — see figure 1.

[3] Chris Harris and Mike Stephens. A Combined Corner and Edge Detector. In *Proceedings of Alvey Conference*, pages 189–192, 1987. 2

[4] Cordelia Schmid and Roger Mohr and Christian Baukhage. Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000. 2

[5] Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two–Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1/2/3):7–42, 2002. 4

[6] David G. Lowe. Distinctive Image Features from Scale–Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004. 2

[7] J. H. Elder. Are edges incomplete ? *International Journal of Computer Vision*, 34:97–122, 1999. 2

[8] J. H. Elder and R. M. Goldberg. Inferential reliability of contour grouping cues in natural images. *Perception*, 27(11), 1998. 3

[9] J. H. Elder and R. M. Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2:324–353, 2002. 1

[10] O. Faugeras. *Three–Dimensional Computer Vision: A Geometric ViewPoint*. MIT Press, 1993. 6

[11] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 41(12), 2001. 2

[12] D. J. Field, A. Hayes, and R. F. Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision Research*, 33(2):173–193, 1993. 3

[13] Frederik Schaffalitzky and Andrew Zisserman. Multi–view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps?". *Lecture Notes in Computer Science*, 2350:414–431, 2002. in Proceedings of the BMVC02. 2

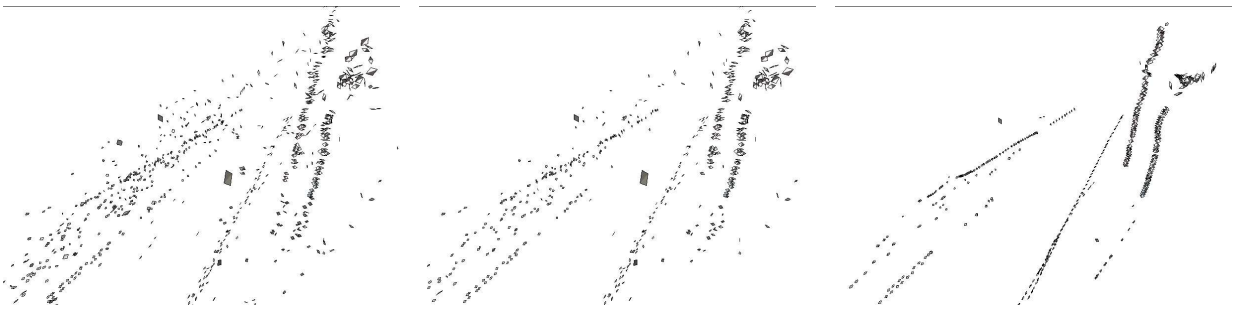
[14] W. Geisler, J. Perry, B. Super, and D. Gallogly. Edge co–occurrence in natural images predicts contour grouping performance. *Vision Research*, 41:711–724, 2001. 3

[15] J. J. Koenderink and A. J. van Doorn. Representation of Local Geometry in the Visual System. *Biological Cybernetics*, 55:367–375, 1987. 2

[16] K. Koffka. *Principles of Gestalt Psychology*. Lund Humphries, London, 1935. 2

[17] K. Köhler. *Gestalt Psychology: An introduction to new concepts in psychology*. New York: Liveright, 1947. 2

[18] N. Krüger. Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, 8(2):117–129, 1998. 3



(a) Stereo reconstruction (b) After external confidence threshold (c) After linear smoothing

Figure 7. Reconstruction of 3D–primitives from stereo–matches obtained from sequence (d) (*c.f.* figure 1). (a) shows the reconstruction resulting from a stereo–matching done using only the multi–modal stereo approach (with a threshold of 0.4 on the multi–modal confidence). (b) shows reconstruction obtained when an additional threshold of 0 is applied to the external confidence. (c) shows the corrected entities, after 3 iterations of the linear smoothing process.

[19] N. Krüger and M. Felsberg. A continuous formulation of intrinsic dimension. In *Proceedings of the British Machine Vision Conference*, 2003. 2

[20] N. Krüger and M. Felsberg. An explicit and compact coding of geometric and structural information applied to stereo matching. *Pattern Recognition Letters*, 25(8), 2004. 1, 2, 4

[21] N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *Interdisciplinary Journal of Artificial Intelligence & the Simulation of Behavious, AISB Journal*, 1(5):417–427, 2004. 1, 2

[22] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, Oct. 2005. 2

[23] Luc Van Gool and Theo Moons and Dorin Ungureanu. Affine / Photometric Invariants for Planar Intensity Patterns. *Lecture Notes In Computer Science*, 1064:642–651, 1996. in Proceedings of the 4th European Conference on Computer Vision — Volume 1. 2

[24] Myron Z. Brown and Darius Burschka and Gregory D. Hager. Advances in Computational Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, Aug. 2003. 4

[25] H.-H. Nagel. On the estimation of optic flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987. 2

[26] P. Parent and S. W. Zucker. Trace interface, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):823–839, 1989. 1

[27] P. Perona and W. Freeman. A factorization approach to grouping. In *Proceedings of the ECCV*, volume 1406, 1998. 1

[28] Peter Kovesi. Image Features from Phase Congruency. *Videre: Journal of Computer Vision Research*, 1(3), 1999. 2

[29] N. Pugeault and N. Krüger. Multi–modal matching applied to stereo. In *Proceedings of the British Machine Vision Conference 2003*, 2003. 4

[30] Ronald Chung and Ramakant Nevatia. Use of Monocular Groupings and Occlusion Analysis in a Hierarchical Stereo System. *Computer Vision and Image Understanding*, 62(3):245–268, Nov. 1995. 1

[31] S. Sarkar and K. L. Boyer. *Computing Perceptual Organization in Computer Vision*. World Scientific Publishing Co. Pte. Ltd., 1994. 1

[32] S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(5):504–525, 2000. 1

[33] A. Sha'ashua and S. Ullman. Grouping contours by iterated pairing network. In *Neural Information Processing Systems (NIPS)*, volume 3, 1990. 1

[34] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 1

[35] M. Wertheimer, editor. *Laws of Organsation in Perceptual Forms*. Harcourt & Brace & Javanowitch, London, 1935. 2

[36] William T. Freeman and Edward H. Adelson. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(9):891–906, Sept. 1991. 2

[37] F. Wörgötter, N. Krüger, N. Pugeault, D. Calow, M. Lappe, K. Pauwels, M. V. Hulle, S. Tan, and A. Johnston. Early cognitive vision: Using gestalt-laws for task-dependent, active image-processing. *Natural Computing*, 3(3):293–321, 2004. 2