OPEN ACCESS Freely available online

PLoS one

Disambiguating Multi–Modal Scene Representations Using Perceptual Grouping Constraints

Nicolas Pugeault[1]*, Florentin Wörgötter[2], Norbert Krüger[3]

1 Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, United Kingdom, **2** Bernstein Center for Computational Neuroscience, Göttingen, Germany, **3** Maersk McKinney Moller Institute, University of Southern Denmark, Odense, Denmark

Abstract

In its early stages, the visual system suffers from a lot of ambiguity and noise that severely limits the performance of early vision algorithms. This article presents feedback mechanisms between early visual processes, such as perceptual grouping, stereopsis and depth reconstruction, that allow the system to reduce this ambiguity and improve early representation of visual information. In the first part, the article proposes a local perceptual grouping algorithm that — in addition to commonly used geometric information — makes use of a novel multi–modal measure between local edge/line features. The grouping information is then used to: 1) disambiguate 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. The integration of mutual feedback between early vision processes is shown to reduce considerably ambiguity and noise without the need for global constraints.

Citation: Pugeault N, Wörgötter F, Krüger N (2010) Disambiguating Multi–Modal Scene Representations Using Perceptual Grouping Constraints. PLoS ONE 5(6): e10663. doi:10.1371/journal.pone.0010663

Editor: Teresa Serrano-Gotarredona, National Microelectronics Center, Spain

Received November 29, 2009; **Accepted** April 20, 2010; **Published** June 9, 2010

Funding: The work described in this paper was funded by the European projects PACOplus and IRFO. Florentin Woergoetter acknowledges funding by the Bernstein Center for Computational Neuroscience, Göttingen. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: n.pugeault@surrey.ac.uk

Introduction

Both human and machine perception involve a progressive abstraction of visual information, from the raw signal provided by the eyes or the cameras towards symbolic, object–centric representations [1]. One problem endemic to visual perception is that each abstraction step requires the taking of some decision about the information, effectively interpreting it; the large amount of noise and ambiguity in the visual signal may lead to erroneous interpretations, as discussed by, e.g., Aloimonos and Shulman [2]. There exist several approaches to solve this problem. One is to design features that describe more closely the original signal, and therefore require less abstraction. However, the resulting representation only describes the appearance of image patches as well as image noise, and lacks a semantic description of shapes — useful, e.g., for grasping, robotic control, planning. Nonetheless, a large amount of work on signal processing and invariant feature descriptors [3] lead to significant progress for tasks like navigation [4] and object recognition [5]. An alternative is to extract abstract symbolic representations directly from the image. One notable attempt by Nevatia and colleagues [6,7], makes use of a feature hierarchy for stereo reconstruction. Another notable class of systems is the model–based vision, where a large amount of world knowledge is available and is used to disambiguate and interpret the visual signal. One problem with the latter approach is that the large amount of ambiguity and noise present in images can lead an early extraction of symbolic features to fail, failures which are difficult to correct. The dilemma between those two approaches can be expressed in terms of the bias/variance dilemma in neural networks [8]. Namely, the use of sophisticated models in vision introduces more bias in the system, whereas signal based approaches lead to more variance.

In the present work, we attempt to address the above dilemma by proposing a gradual abstraction that postpones decision taking using mutual feedback between two mid–level visual processes, namely perceptual grouping and stereopsis, to reduce ambiguity and noise. Ambiguities addressed here include incorrect stereo matches and inaccurate 3D reconstructions. Moreover, properties of the local signal such as local estimates of orientation, phase and colour will also be stabilised by perceptual grouping mechanisms. This work makes use of a sparse symbolic scene representation based on multi–modal *primitives* [9]. In this work, the term 'multi–modal' stresses that the descriptors cover different *visual* modalities such as motion, orientation and colour; it is not meant to indicate different *sensorial* modalities. Primitives form a local feature vector containing multi–modal visual information covering appearance as well as geometric information, in 2D and 3D. Such multi–modal descriptors offer certain advantages for the representation of visual scenes. For example, they allow for the explicit formulation of visual semantics in terms of meaningful local descriptors and higher–order relations between them, such as motion, co–planarity and similarity of appearance (see, e.g., [10]). One property of symbolic representations is that the transfer of visual information to a symbolic level increases the predictiveness of visual events [11] and at the same time decreases the memory and bandwidth required to process and transfer information. Hence, in these representations, regularities between visual events can be efficiently used for disambiguation. Primitives–based visual representations are used in a variety of

applications, covering, e.g., object learning [12] and grasping [13].

The contributions in this paper are threefold: first we propose a local perceptual grouping mechanism making full use of the multi–modal and semantic information carried by the visual primitives; second, we propose a stereo matching scheme for primitives, allowing for the reconstruction of the 3D equivalent of 2D primitives; third, we investigate how perceptual grouping reduces ambiguities in the reconstructed 3D representation. In the following, these contributions will be described in more detail and put into the context of related work.

This paper's first contribution is a perceptual grouping scheme making use of the multi–modal information carried by the primitives. Perceptual grouping can be divided into two tasks: 1) defining an affinity measure between primitives and using it to build a graph of the connectedness between 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 a primitive's local grouping information to apply the correction mechanisms proposed in this paper. Similar affinity measures have been proposed [14,15], formalising a *good continuation* constraint, and Elder and Goldberg [16] included the intensity on each side of the contour into a Bayesian formulation of grouping. We go beyond this work by proposing 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.

As a second contribution, this work extends the work by Krueger and Felsberg [17] by enriching the multi–modal stereo matching using local motion [18] and, more importantly, by evaluating statistically the importance of the different visual modalities for stereo matching using ground truth range data.

As a third contribution, we make use of perceptual groups of primitives to disambiguate stereo matching and correct the 3D scene reconstruction. Grouping allows for the interpolation of visual properties such as position, local orientation, phase and colour, and thus helps to improve local feature extraction. This paper studies how perceptual grouping information can be used to disambiguate stereopsis and 3D reconstruction using primitives. If we assume that image contours (2D) are likely to be the projection of 3D contours on the image, then we can expect all 3D contours to project as 2D contours on each camera plane (except in the case of partial occlusions). Conversely, this also implies that any contour in one image has a corresponding contour in the second image. We therefore propose a non–local *external* stereo confidence measure, which estimates how well a primitive's neighbours that belong to the same group agree with that primitive's putative stereo correspondences. This allows for discarding a large number of putative stereo correspondences, hence reducing the ambiguity of the stereo matching and scene reconstruction processes. Moreover, the interpolation of the curves described by groups of primitives is used to correct these primitives' geometric and appearance modalities.

The scheme presented in this paper is illustrated in Figure 1, where solid lines stand for forward dependencies and dashed lines for feedback mechanisms. The local symbolic representation is extracted from the images. From this representation, we extract perceptual groups (i.e., contours) and we use correspondences across a pair of stereo views of the scene to reconstruct a local and symbolic 3D representation of the scene, equivalent to the 2D image representations it is reconstructed from; this is the feedforward part of the scheme, represented with solid lines. Then, the perceptual grouping information is used to correct the 2D symbolic image description, the stereo matches, and the reconstructed 3D scene representation; this is the corrective part of the scheme, represented with dashed lines.

Methods

This section is structured as follows: first, the multi–modal primitives are described; second, distance measures for all modalities are proposed; third, the grouping mechanism is presented; fourth, the stereo matching scheme is discussed; then, a scheme for increasing stereo matching reliability from grouping information is described; finally, we present a scheme to correct 2D and 3D primitives' position and orientation by interpolating the curves described by groups of primitives.

2D primitives

Numerous feature detectors exist in the literature (see Mikolajczyk and Schmid [3] for a review). Any feature based approach can be divided into two complementary tasks: an interest point detector [19,20] and a descriptor encoding information from a local patch of the image at this location, that can be based on histograms [3,21], spatial frequency [22–24], local derivatives [25–27], steerable filters [28], or invariant moments [29]. In [3], these different descriptors have been compared, showing a best performance for SIFT–like descriptors (Scale Invariant Feature Transform [21]).

The primitives we will use in this work are local, multi–modal edge descriptors, described in Ref. [9]. In contrast to the above mentioned features, 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 purpose are discussed by Elder [30].

In the first step, an event map and the associated local phase are computed using the *monogenic signal* [31] — note that other signal processing could alternatively be used (e.g., steerable filters [28]). The 2D primitives are sparsely extracted at locations in the image that are most likely to contain events (edges or lines); these locations are detected using the local intrinsic dimension [32]. Sparseness is assured using a classical winner–take–all operation, which guarantees that the extracted primitives describe different image patches. Multi–modal information is gathered locally from the image, including the position $\boldsymbol{x}$ of the centre of the patch, the orientation θ of the event, the phase ϕ of the signal at this point, the colour $\boldsymbol{c}$ sampled over the image patch on both sides of the event, and the local optical flow $\boldsymbol{f}$ computed using the classical Nagel algorithm [33] (the flow is disregarded for still images). The phase encodes the type of contrast transition across the event, e.g., dark to bright edge or dark line on bright background. See Ref. [22–24]. Consequently, a primitive is described by the multi–modal vector

$$\boldsymbol{\pi} = (\boldsymbol{x}, \theta, \phi, \boldsymbol{c}, \boldsymbol{f})^T. \tag{1}$$

The set of primitives describing an image is called *image representation* and written $\mathcal{I}^l$ and $\mathcal{I}^r$ for images from the left and right camera. The image representation extracted from one image is illustrated in Figure 2. In the upper–left corner, panel A shows one image extracted from an indoor video sequence; panel B

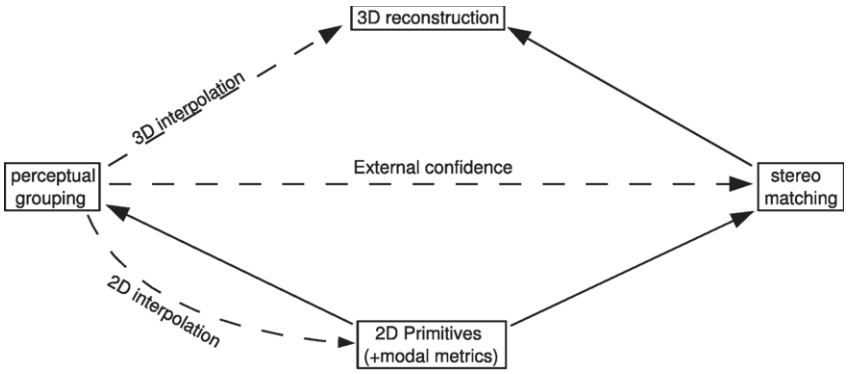

Figure 1. Summary of the scheme presented in this paper. In this figure, solid arrows mean direct dependencies and dashed lines corrective feedback.
doi:10.1371/journal.pone.0010663.g001

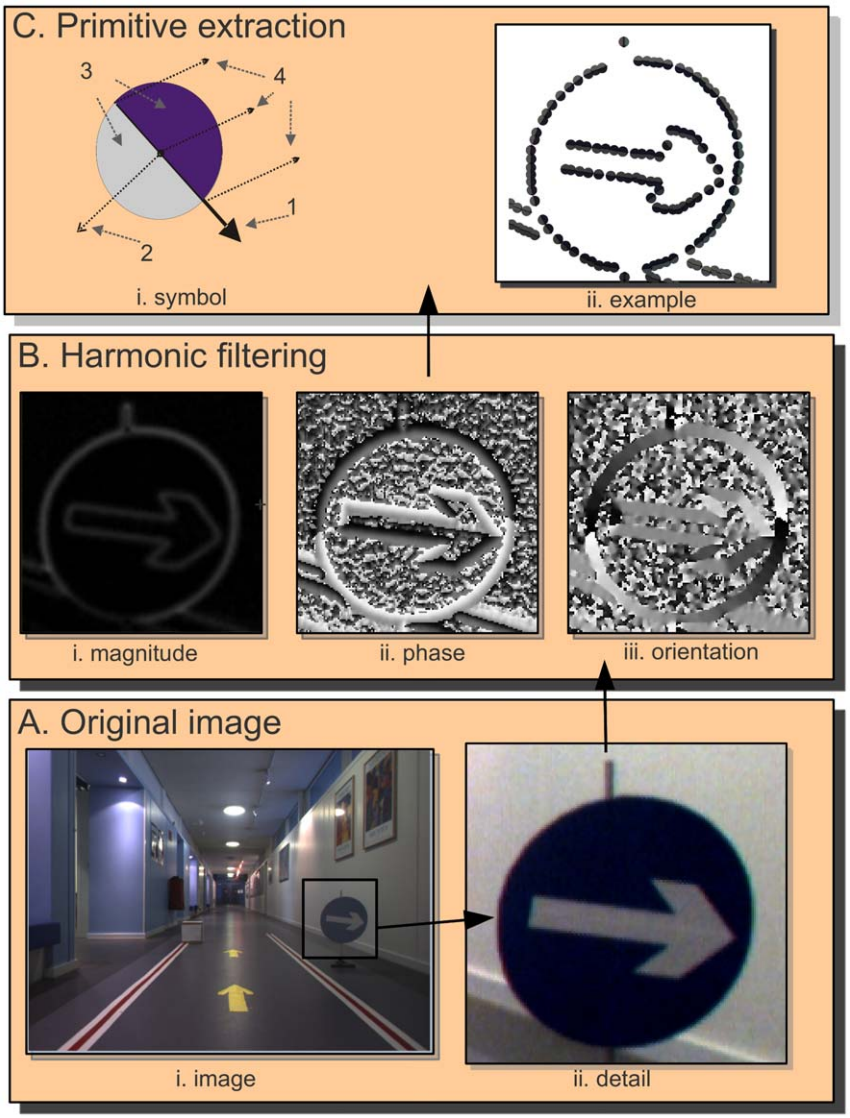

Figure 2. Illustration of the primitive extraction process from an indoor video sequence. A The original image and a magnified detail. **B** Harmonic filtering (using, e.g., Gabor wavelets, monogenic signal or steerable filters) provides estimates of the local (i) magnitude, (ii) orientation, and (iii) phase of the signal. **C** Primitive extraction: (i) the symbolic primitive, where 1 stands for the orientation, 2 for the phase, 3 for the colour, and 4 for the optic flow; (ii) example of the primitives extracted from the image detail.
doi:10.1371/journal.pone.0010663.g002

shows the result of a local filtering; and panel C shows the extracted primitives.

Note that these primitives are of lower dimensionality than, e.g., SIFT features (12 vs. 128) and can therefore suffer from a lesser distinctiveness (two unrelated primitives have a greater chance to have a similar aspect). Nonetheless, we will show in the results section that they are distinctive enough for a reliable stereo matching if the epipolar geometry of the cameras is known. The rich information carried by the 2D primitives can be used to reconstruct them in 3D, providing a more complete scene representation. Geometric meaning allows a description of proximate primitives in terms of perceptual grouping, as will be discussed in the following section.

Metrics of 2D primitives

In this section, we define metrics for each of the primitives' modalities. Those metrics will be used in the following for perceptual grouping of primitives and for stereo matching. Figure 3 illustrates how the distance measures defined here are combined. In the case of perceptual grouping (solid lines), proximity, collinearity and co–circularity measures between a pair of primitives are merged into a Geometric affinity, whereas the distances in phase, colour and optic flow form the Multi–modal affinity. The combination of those two form the overall affinity $c[g_{i,j}]$ that is used to group 2D primitives. In the case of stereopsis (dashed lines) the orientation distance between the two primitives replaces the geometric criterion. Then the multi–modal similarity is computed from orientation, phase, colour and optic flow distances.

Note that, in the context of perceptual grouping, the orientation difference is replaced with a more sensible interpretation of the good continuation constraint, combining proximity, collinearity and co–circularity; in contrast, the stereo similarity makes direct use of the orientation difference.

Orientation: If we consider two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$, respectively with the orientations θ_i and θ_j, then their orientation distance is

$$d_\theta(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\frac{2}{\pi}\left|\arctan\left(\sin(\theta_j-\theta_i),\cos(\theta_j-\theta_i)\right)\right|. \qquad (2)$$

The $\frac{2}{\pi}$ factor ensures that the orientation metric is between [0,1], with 0 standing for parallel orientations, 0.5 for a 45 degrees angle and 1 for orthogonal orientations.

Phase: The phase metric d_ϕ is

$$d_\vartheta(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\frac{1}{\pi}\left|\arctan\left(\sin(\phi_j-\phi_i),\cos(\phi_j-\phi_i)\right)\right|. \qquad (3)$$

The $\frac{1}{\pi}$ factor ensures that the phase metric is between [0,1], with 0 standing for two primitives encoding the contrast transition (e.g., bright to dark edge), and 1 standing for opposite contrast (e.g., a dark line and a bright line).

Colour: The colour metric d_c is

$$d_c(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\frac{1}{2}\sum_{q\in\{l,r\}} d_{c,q}, \qquad (4)$$

where $d_{c,q}$ is defined in HSV space as

$$d_{c,q}(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\begin{cases}\dfrac{d_\alpha(H_i^q,H_j^q)+|S_j^q-S_i^q|+|V_j^q-V_i^q|}{3} & \text{if } V>0.1,S>0.1,\\ \dfrac{|S_j^q-S_i^q|+|V_j^q-V_i^q|}{2} & \text{if } V>0.1,S\le 0.1,\\ |V_j^q-V_i^q| & \text{otherwise.}\end{cases} \qquad (5)$$

Because of the conical topology of the HSV space, the hue component H is basically random for very low saturation S, and saturation is random for low values of V. This equation discards hue information for low saturation, and saturation information for low value of V, and otherwise weights evenly the colour components. In Eq. 5, d_α stands for the angular distance

$$d_\alpha(\alpha_1,\alpha_2)=\frac{1}{\pi}\left|\arctan(\sin(\alpha_2-\alpha_1),\cos(\alpha_2-\alpha_1))\right|, \qquad (6)$$

and H_i^l (H_i^r), S_i^l (S_i^r) and V_i^l (V_i^r) are the hue, saturation and value components on the left (right) side of the primitive $\boldsymbol{\pi}_i$.

Optic Flow: The optic flow d_f metric is

$$d_f=\frac{1}{\pi}\arccos\left(\frac{\boldsymbol{f}_i\cdot\boldsymbol{f}_j}{\max(\|\boldsymbol{f}_i\|,\|\boldsymbol{f}_j\|)}\right). \qquad (7)$$

Note that these metrics are the same used in Refs. [17,18].

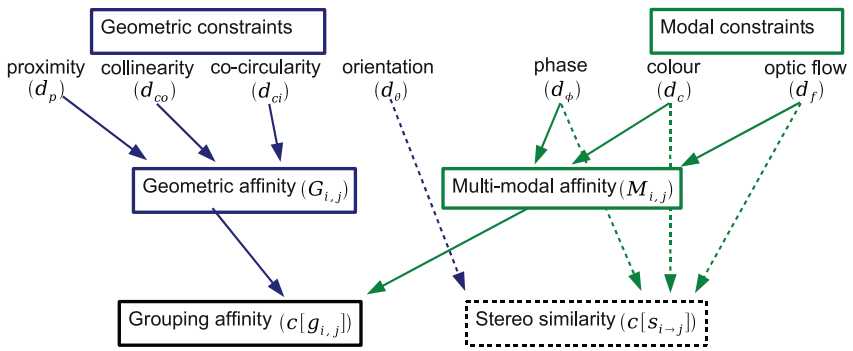

Figure 3. Illustration of the measures used in this paper and how they are combined. Solid arrows indicate the metrics used for stereopsis, dashed lines the metrics used for perceptual grouping.
doi:10.1371/journal.pone.0010663.g003

Perceptual grouping of 2D primitives

Since the 1930's, the Gestalt psychologists suggested a collection of axioms describing the way the human visual system binds together features in an image [34–36]. This process is generally called *perceptual grouping* and the Gestalt psychologists proposed that it is driven by properties like proximity, good continuation, similarity and symmetry, amongst others. More recently, psychophysical experiments measured the impact of different cues for perceptual grouping (see, e.g., Ref. [37]). Furthermore, Brunswik and Kamiya [38] postulated that these properties should be related to statistics of natural images. This was later confirmed by several studies [39–41].

We defined the primitives as local edge descriptors, and assumed that a group of primitives describes a contour in 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*, image contours are expected to be continuous and smooth (small and constant local curvature); thus, two proximate primitives in a group are expected to be either nearly collinear, or co–circular. According to these rules, 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 of the contour it describes. Finally, we would expect a contour's properties such as colour (on both sides) to change smoothly (or not at all) along this contour. This is formalised by the rule of *similarity*, which states that similar primitives (in terms of the colour, phase and optical flow modalities) are most likely to belong together.

The two first rules are joined into a *Geometric constraint*, that is combined with a multi–modal *Appearance constraint* into an overall affinity measure.

Geometric constraints. The first constraint we enforce during grouping stems directly from the symbolic quality of the primitives: primitives are local event descriptors and therefore, according to the good continuation law, they should be locally nearly collinear or co–circular to form a group. Effectively, we compute this constraint as a combination of proximity, collinearity and co–circularity measures.

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 $\mathcal{C}$ can be formulated as a combination of three basic constraints on their relative position and orientation — see Figure 4.

Proximity: The proximity measure is given by

$$d_p(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\exp\left[-\max\left(1-\frac{\|\mathbf{v_{ij}}\|}{\rho\mu},0\right)\right]. \quad (8)$$

Here, ρ stands for the radius of the primitive in pixels, and the quantity $\rho\mu$ is the maximal distance between two primitives for them to be compared; more distant primitives will not be compared and therefore have a null similarity. The quantity $\|\mathbf{v_{ij}}\|$ stands for the distance (in pixels) separating the two primitives' centres. We found experimentally that $\mu=5$ proved to be a good value — i.e., grouped primitives are distant by five timed their size at most.

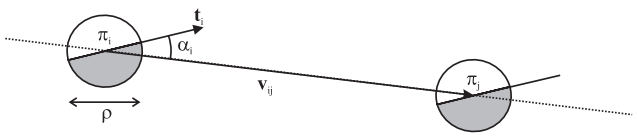

Figure 4. Illustration of the values used for the collinearity computation. If we consider two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$, then the vector between the centres of these two primitives is written $\mathbf{v_{ij}}$, and the orientations of the two primitives are designated by the vectors $\mathbf{t}_i$ and $\mathbf{t}_j$, respectively. The angle formed by $\mathbf{v_{ij}}$ and $\mathbf{t}_i$ is written α_i, and between $\mathbf{v_{ij}}$ and $\mathbf{t}_i$ is written α_j. ρ is the diameter of the primitive in pixels.
doi:10.1371/journal.pone.0010663.g004

Collinearity: The collinearity measure is

$$d_{co}(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\left|\sin\left(\frac{|\alpha_i|+|\alpha_j|}{2}\right)\right|. \quad (9)$$

Co–circularity: The co–circularity measure is

$$d_{ci}(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)=\left|\sin\left(\frac{\alpha_i+\alpha_j}{2}\right)\right|, \quad (10)$$

where α_i and α_j are the angles between the line joining the two primitives centres and the orientation of $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$, respectively (see figure 4).

Geometric affinity: The combination of those three criteria forms the geometric constraint:

$$G_{i,j}=\left(\prod_{x\in\{p,co,ci\}}\left(1-d_x(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)\right)\right)^{\frac{1}{3}} \quad (11)$$

where $G_{i,j}$ is the geometric affinity between two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$. This affinity models the likelihood of a curve tangent to the lines defined by the two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_i$; we have $G_{i,j}=1$ for a perfect match.

Appearance constraints. Effectively, the more similar the modalities between two primitives are, the more likely are those two primitives part of the same event. Note that Elder and Goldberg [39] already proposed to use the intensity as a cue for perceptual grouping, yet here we use a combination of phase, colour, and optical flow modalities of the primitives to decide, using the value of $\mathbf{M}$, if they describe the same event.

Appearance affinity: The appearance–based affinity is

$$M_{i,j}=1-\sum_{m\in\{\phi,\boldsymbol{c},\boldsymbol{f}\}}w_m d_m\left(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j\right), \quad (12)$$

where w_m is the relative weighting of the modality $m\in\{\phi,\boldsymbol{c},\boldsymbol{f}\}$, with $\sum_{m\in\{\phi,\boldsymbol{c},\boldsymbol{f}\}}w_m=1$, and d_m refers to the metrics defined in equations 3, 4, and 7; the modality weights were all set to $w_m=\frac{1}{3}$; Therefore, $M_{i,j}=1$ stands for a perfect match between two primitives. Because the geometric constraint models the relative orientation of two primitives in a manner more adapted to the problem of grouping line segments, the orientation metric is not part of the multi–modal constraint.

Overall affinity. We define this affinity from Equations (11) and (12), 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 with strongly dissimilar modalities, will only have an average affinity.

Two primitives $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ form a *link* $g_{i,j}$ if they share a significant affinity (significant being set by a threshold on the overall affinity), and the confidence $c\left[g_{i,j}\right]$ of this link is given by

the overall affinity:

$$c[g_{i,j}] = \sqrt{G_{i,j} \cdot M_{i,j}}. \quad (13)$$

We found experimentally that applying a threshold of $c[g_{i,j}] \geq 0.5$ yields a good grouping, as can be seen in Figure 5.

This affinity is also a valid estimate of the likelihood for $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ to be part of the same contour $\mathcal{C}$. In the following, we will consider that a link $g_{i,j}$ between two primitives exists if its confidence $c[g_{i,j}]$ is large enough. We will call *neighbourhood* $\mathcal{N}(\boldsymbol{\pi}_i)$ of a primitive $\boldsymbol{\pi}_i$ all primitives $\boldsymbol{\pi}_j$ such that $g_{i,j}$ is a link:

$$\mathcal{N}(\boldsymbol{\pi}_i) = \{\boldsymbol{\pi}_j | \exists g_{i,j}\}. \quad (14)$$

Figure 6 shows the links extracted, along with the different modal affinities. The links extracted for different thresholds τ_A on the affinity are shown in Figure 5. In the following, links are extracted only if $c[g_{i,j}] > 0.5$. The lines in these figures describe strings of grouped primitives. One can see in these images that the major image contours are adequately described. This criterion is what is meant in the rest of the paper every time we refer to 'groups'.

Stereopsis using 2D primitives

In this section, we extend the concept of multi–modal primitives to 3D: first, we define a local multi–modal matching function; then we define the 3D primitives.

Classical stereopsis [42,43] allows for the reconstruction of 3D points from pairs of corresponding points in two stereo images. A review of stereo algorithms was presented by Brown et al. [44].

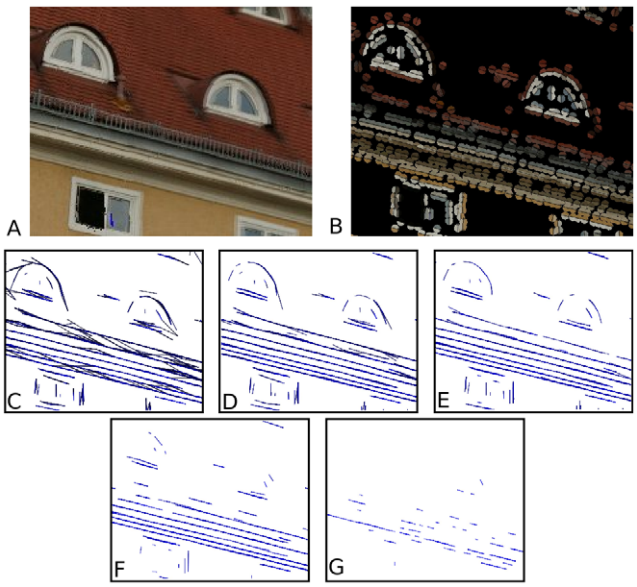

Figure 5. Illustration of the links extracted for different affinity thresholds. A detail of the original image (220×280 pixels); **B** extracted primitives; **C**–**G**. extracted links for values of $\tau_A =$ (**C**) 0.1, (**D**) 0.3, (**E**) 0.5, (**F**) 0.7, and (**G**) 0.9 — using $\mu = 5$. The blue lines represent the links, where more saturated lines stand for higher affinity values.
doi:10.1371/journal.pone.0010663.g005

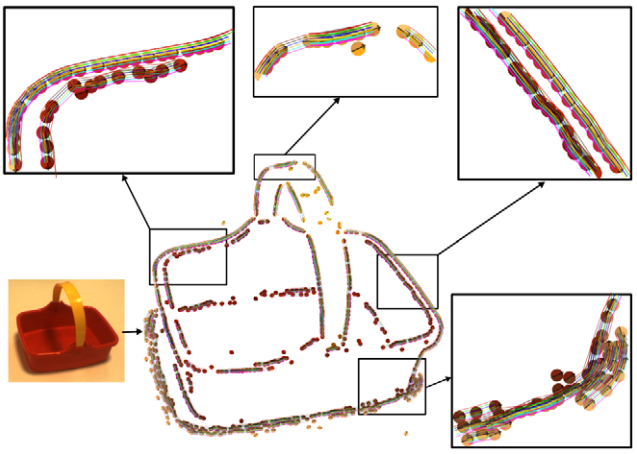

Figure 6. Illustration of the affinities between 2D primitives. In this figure, the 2D primitives are linked by coloured lines, where a brighter colour stands for a stronger affinity. Red stands for collinearity, green for phase, blue for colour and yellow for optical flow affinity.
doi:10.1371/journal.pone.0010663.g006

Dense two–frames stereo algorithms (i.e., matching each and every pixel in the first image with a pixel in the second) were also compared by Scharstein and Szeliski [45]. The present work differs from classical approaches insofar that symbolic multi–modal entities are matched, and reconstructed, rather than points. Although it is commonplace to use complex features (e.g., SIFT) for matching, only the locations in space are generally reconstructed, whereas the present work reconstructs a symbolic local interpretation in space. The proposed method is local and makes use of the epipolar constraint to limit the scope of the correspondence search.

If we consider a 2D primitive $\boldsymbol{\pi}_i$ in the left image $\mathcal{I}^l$, all 2D primitives $\boldsymbol{\pi}_p$ in the right image that lie nearby its epipolar line ξ_i are considered as *putative correspondences*, written $s_{i \to p}$. The difference between the image coordinates of $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_p$ is generally called the *disparity*. We will differentiate between the orthogonal distance from the centre of $\boldsymbol{\pi}_p$ to the epipolar line ξ_i, called *normal disparity*, and the distance along this line, called *tangential disparity*. The normal disparity expresses how strictly the epipolar constraint is satisfied. A certain tolerance is required here due to the representation's sparseness. In the following all primitives with a normal disparity lower than 1.5 times the primitives' size are considered. The tangential disparity has a direct relation with the depth of the reconstructed 3D primitive: a tangential disparity of zero means that the point is infinitely far, whereas larger disparities denote closer points.

Finally, one putative correspondence $s_{i \to p}$ is chosen using a local winner–take–all scheme: all putative correspondences $\boldsymbol{\pi}_p \in \mathcal{I}^r$ (in the right image) of a primitive $\boldsymbol{\pi}_i \in \mathcal{I}^l$ (in the left image) are competing against each other. The confidence in each of them is set to their *similarity* with the left primitive $\boldsymbol{\pi}_i$, and the most similar correspondence is selected. This similarity measure is explained in the following section.

Multi–modal stereo similarity. The multi–modal distance between two primitives is defined as a linear combination of the modal distances between two primitives. This similarity is akin to the multi–modal affinity defined in Equation (12) with the addition of the orientation similarity, that is used here to replace the geometric constraint:

$$c\left[s_{i\to j}\right]=1-\sum_{m\in\{\theta,\phi,\boldsymbol{c},\boldsymbol{f}\}} w_m d_m(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j), \tag{15}$$

where w_m is the relative weighting of the modality $m\in\{\theta,\phi,\boldsymbol{c},\boldsymbol{f}\}$, with $w_m\geq 0$ and $\sum_{m\in\{\theta,\phi,\boldsymbol{c},\boldsymbol{f}\}} w_m=1$. The performance of a winner–take–all stereo matching scheme based on this multi–modal similarity is evaluated on several stereo sequences in the results section.

Reconstruction of 3D primitives. We propose to reconstruct the 3D equivalent of a stereo pair of corresponding 2D primitives, hereafter called *3D primitives* ($\boldsymbol{\Pi}$) as encoded in the vector:

$$\boldsymbol{\Pi}=(\boldsymbol{X},\boldsymbol{\Theta},\Phi,\boldsymbol{C})^{\mathrm{T}} \tag{16}$$

where $\boldsymbol{X}$ is the location in space, $\boldsymbol{\Theta}$ is the 3D orientation of the edge, Φ is the phase across this edge, and $\boldsymbol{C}$ holds the local colour information on both sides of the contour. Figure 7 illustrates the reconstruction of a 3D primitive from a stereo pair of corresponding 2D primitives. A 2D primitive defines an image line, that back–projects as a 3D plane; the intersection between the two planes back–projected by the corresponding primitives provide a 3D line, onto which the 3D primitive lies. This line's orientation give the 3D primitive's orientation; its position is given by the intersection between the line back–projected by the first 2D primitive's position, and the plane back–projected by the corresponding 3D primitive. We refer to [46] for a complete discussion of the 3D primitives reconstruction.

The reconstruction shown corresponds to a multi–modal winner–take–all matching (using equation (15)) with a similarity threshold set to $\tau_m=0.5$.

Perceptual grouping of 3D primitives. In order to allow for reasoning in the 3D space, we extend the perceptual grouping defined for 2D primitives to the reconstructed 3D primitives.

Two 3D primitives $\boldsymbol{\Pi}_i$ and $\boldsymbol{\Pi}_j$ are linked $g^{3D}_{i,j}$, if and only if their projection in both image planes (respectively $\boldsymbol{\pi}^l_i$ and $\boldsymbol{\pi}^l_j$ on the left image and $\boldsymbol{\pi}^r_i$ and $\boldsymbol{\pi}^r_i$ on the right) are linked (such that the two links $g^l_{i,j}$ and $g^r_{i,j}$ both exist), according to the logical implication

$$g^l_{i,j}\wedge g^r_{i,j}\Rightarrow g^{3D}_{i,j}. \tag{17}$$

This definition extends naturally the perceptual groups defined in the image domain to the 3D space.

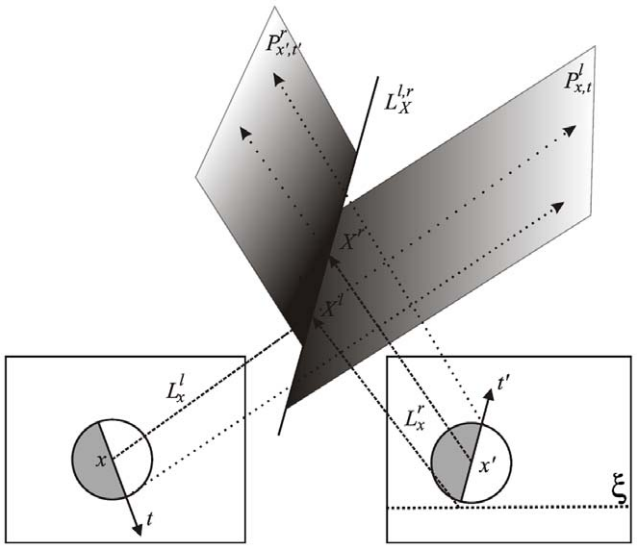

Figure 7. Illustration of a 3D primitive reconstruction from a stereo pair of 2D primitives. Each 2D primitive defines an image line, that back–projects as a plane in 3D space. The intersection of these two 3D planes yield a line in space that defines the 3D primitive's orientation. The 3D primitive's position is given by the intersection between the back–projections of both 2D primitives' position. We refer to [46] for a complete discussion of the 3D primitives reconstruction.
doi:10.1371/journal.pone.0010663.g007

Perceptual grouping constraints to improve stereopsis

In this section, we define a semi-global stereo matching function that is based on the expected consistency between grouping processes in the left and right image as well as the stereo matching process. We show that matching can be improved significantly by using such kind of context information. It also allows for the establishment of groups in 3D for which additional interpolation processes can be applied to further improve the precision of reconstruction.

Because the primitive–based image representation used in this work samples lines and step–edges, it carries redundant information along contours. This redundancy can be used for constraining the stereo matching problem, leading to the two following constraints:

(C1) Isolated primitives are likely to be unreliable: As primitives are extracted redundantly along the contours, conversely an isolated primitive is likely to be an artefact and hence isolated primitives can be neglected.

(C2) 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 (notwithstanding pathological cases).

In our representation, contour information is encoded by the link network that is the result of the perceptual grouping mechanism presented earlier; this is illustrated in Figure 8. In this figure, the orientation of the primitive $\boldsymbol{\pi}_i$ makes it the most similar (according to Equation (15)) to $\boldsymbol{\pi}_2$; hence, the stereo correspondence $s_{2\to i}$ holds a higher confidence than, e.g., $s_{2\to j}$. However, the putative correspondence $\boldsymbol{\pi}_j$ forms a group $g_{s,j}$, thus preserving the group relation $g_{1,2}$ across stereo, whereas $\boldsymbol{\pi}_i$ is not grouped with $\boldsymbol{\pi}_s$. Therefore, $\boldsymbol{\pi}_j$ is more likely to be the true stereo correspondence of $\boldsymbol{\pi}_2$.

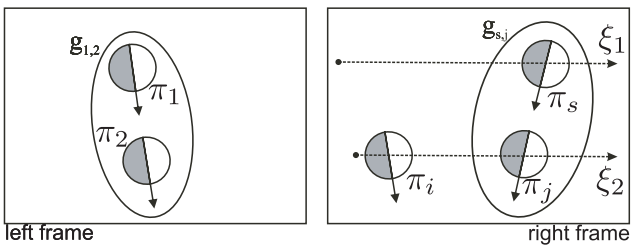

Figure 8. The BSCE criterion. Let $\boldsymbol{\pi}_1$ be a primitive in the left frame forming a group with a second primitive $\boldsymbol{\pi}_2$. $\boldsymbol{\pi}_1$ has a stereo correspondence $\boldsymbol{\pi}_s$ that lie on the epipolar line ξ_1 in the right image. Both $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ in the right image lie on the epipolar line ξ_2 of $\boldsymbol{\pi}_2$; hence these two primitives are both putative correspondences of $\boldsymbol{\pi}_2$.
doi:10.1371/journal.pone.0010663.g008

Basic Stereo Consistency Event (BSCE). Primitives represent local estimators of image contours; a constellation of primitives describes a contour as a whole. Such contours are consistent over stereo, with the notable exception of occlusion cases. As we have defined the likelihood for two primitives to describe the same contour as the affinity between these two primitives, we can rewrite the previous statement as:

Definition 1 *Given two primitives $\boldsymbol{\pi}_i^l$ and $\boldsymbol{\pi}_j^l$ in the left image $\mathcal{I}^l$ and their respective correspondences $\boldsymbol{\pi}_n^r$ and $\boldsymbol{\pi}_p^r$ in the right image $\mathcal{I}^r$; if $\boldsymbol{\pi}_i^l$ and $\boldsymbol{\pi}_j^l$ belong to the same group in $\mathcal{I}^l$, then $\boldsymbol{\pi}_n^r$ and $\boldsymbol{\pi}_p^r$ should also be part of a group in $\mathcal{I}^r$.*

The link conservation between a pair of primitives and the stereo correspondences thereof is called *Basic Stereo Consistency Event* (BSCE) [47]. This condition can then be used to test the validity of a stereo hypothesis. Consider a primitive $\boldsymbol{\pi}_i^l$, a stereo hypothesis

$$s_{i\to n} : \boldsymbol{\pi}_i^l \to \boldsymbol{\pi}_n^r, \tag{18}$$

and a 2D primitive $\boldsymbol{\pi}_j^l \in \mathcal{N}(\boldsymbol{\pi}_i^l)$ in the neighbourhood of $\boldsymbol{\pi}_i^l$ (as defined in Equation (14)), such that the two primitives share an affinity $c\left[g_{i,j}\right]$ — see Equation (13). For this second primitive, a stereo correspondence $\boldsymbol{\pi}_p^r$ with a confidence of $c\left[s_{j\to p}\right]$ exists. We can now define an estimate of how well the stereo hypothesis $s_{i\to n}$ reflects the BSCE by:

$$E(g_{i,j},s_{i\to n}) = \begin{cases} +\sqrt{c\left[s_{j\to p}\right]\cdot c\left[g_{i,j}\right]} & \text{if } c\left[g_{n,p}\right] > \tau_A \\ -\sqrt{c\left[s_{j\to p}\right]\cdot c\left[g_{i,j}\right]} & \text{otherwise} \end{cases}. \tag{19}$$

In other words: the BSCE between a primitive in the first image and one of its neighbours is high if they share a strong affinity and if both primitives' stereo correspondences in the second image *also* share a strong affinity; it is low if they share a strong affinity yet their stereo correspondences in the second image do not. This naturally extends the concept of group into the stereo domain.

Neighbourhood consistency confidence. Equation (19) tells us how a primitive's stereo correspondence is consistent with our knowledge of one of its neighbours' stereo correspondence. In this section we extend this definition to the whole primitive's neighbourhood. 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 $\mathcal{N}(\boldsymbol{\pi}_i^l)$ — as defined by Equation (14) —

$$c_{ext}[s_{i\to n}] = \frac{1}{\#\mathcal{N}(\boldsymbol{\pi}_i^l)} \sum_{\boldsymbol{\pi}_k^l \in \mathcal{N}(\boldsymbol{\pi}_i^l)} E(g_{i,k},s_{i\to n}), \tag{20}$$

where $\#\mathcal{N}(\boldsymbol{\pi}_i^l)$ 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 primitives — Equation (15).

Correcting primitives using contextual knowledge

Although primitives are extracted with sub–pixel localisation, their actual accuracies vary to a large extent depending on local amounts of noise, blur and texture in the image. The primitives' position and orientation inaccuracy is amplified by stereo reconstruction [48] and can lead to large errors thereafter. Moreover, one fundamental drawback of stereo–based reconstruction of 3D shapes is that the reconstructed entities' precision decreases quickly with distance to the cameras, due to the images' finite pixel sampling [49,50]. The symbolic quality of primitives, and groups of primitives, provides us with additional knowledge that can be used to reduce this uncertainty. Namely, groups of 3D primitives are reconstructed from pairs of 2D primitives that form a perceptual group in both stereo images, and as such, according to the grouping assumption, they describe a smooth and continuous contour of the scene (except in some pathological perspectives). This knowledge that the group as a whole should form a smooth contour can be used to correct the individual 3D primitives modalities. In this section, we propose a scheme for correcting 2D– and 3D primitives by locally interpolating the contours described by groups of primitives.

Triplets of primitives. If we consider three primitives $\boldsymbol{\pi}_i$, $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$, which belong to the same group, and if $\boldsymbol{\pi}_i$ lies in between $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$ — such that the Euclidean distances between $(\boldsymbol{\pi}_i,\boldsymbol{\pi}_j)$ and $(\boldsymbol{\pi}_i,\boldsymbol{\pi}_k)$ are both smaller than that between $(\boldsymbol{\pi}_j,\boldsymbol{\pi}_k)$ — then we call $\mathbf{t}_{ijk} = (\boldsymbol{\pi}_i,\boldsymbol{\pi}_j,\boldsymbol{\pi}_k)$ a *triplet*. Formally,

$$g_{i,j} \wedge g_{i,k} \wedge \left(\max\left(\|\boldsymbol{x}_j - \boldsymbol{x}_i\|, \|\boldsymbol{x}_k - \boldsymbol{x}_i\|\right) < \|\boldsymbol{x}_k - \boldsymbol{x}_j\|\right) \Rightarrow \mathbf{t}_{ijk}. \tag{21}$$

Triplets of 3D primitives can be defined in the exact same manner in 3D space: as for the 2D case, a 3D triplet $\mathbf{t}_{ijk}^{3D} = (\boldsymbol{\Pi}_i,\boldsymbol{\Pi}_j,\boldsymbol{\Pi}_k)$ is constituted of a central primitive $\boldsymbol{\Pi}_i$ linked to two supporting primitives $\boldsymbol{\Pi}_j$ and $\boldsymbol{\Pi}_k$, such that the central primitive lies in between the two supporting primitives (i.e., the Euclidean distances between $(\boldsymbol{\Pi}_i,\boldsymbol{\Pi}_j)$ and $(\boldsymbol{\Pi}_i,\boldsymbol{\Pi}_k)$ are both smaller than $(\boldsymbol{\Pi}_j,\boldsymbol{\Pi}_k)$). Formally,

$$g_{i,j}^{3D} \wedge g_{i,k}^{3D} \wedge \left(\max\left(\|\boldsymbol{X}_j - \boldsymbol{X}_i\|, \|\boldsymbol{X}_k - \boldsymbol{X}_i\|\right) < \|\boldsymbol{X}_k - \boldsymbol{X}_j\|\right) \Rightarrow \mathbf{t}_{ijk}^{3D}. \tag{22}$$

These triplets are useful because it is possible to interpolate the curve between two primitives, and therefore, we can use the curve interpolated between the two supporting primitives of the triplet ($\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$) to correct the central primitive ($\boldsymbol{\pi}_i$).

Interpolation of modalities. We interpolate the curve between two (2D or 3D) primitives using Hermite polynomials [51]. These are convenient in this context as they allow for the interpolation of a curve from only two data points and the curve tangents at those points. Also, Hermite splines can be applied to interpolate 2D or 3D curves indifferently.

Position and orientation: The curve interpolated between two primitives $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$, with positions $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$, and local tangents (defined by the primitives' orientations) of $\boldsymbol{t}_j$ and $\boldsymbol{t}_k$ is defined as all the points $\hat{\boldsymbol{x}}_s$ in the image, with $s \in [0,1]$ such that $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_j$ and $\hat{\boldsymbol{x}}_1 = \boldsymbol{x}_k$ and

$$\hat{\boldsymbol{x}}_s = \begin{pmatrix} s^3 \\ s^2 \\ s \\ 1 \end{pmatrix} \cdot \boldsymbol{H} \cdot \begin{pmatrix} \boldsymbol{x}_j \\ \boldsymbol{x}_k \\ \boldsymbol{t}_j \\ \boldsymbol{t}_k \end{pmatrix}, \tag{23}$$

where $\boldsymbol{H}$ is the matrix formulation for the Hermite polynomials

$$\boldsymbol{H}=\begin{pmatrix}2 & -2 & 1 & 1\\ -3 & 3 & -2 & -1\\ 0 & 0 & 1 & 0\\ 1 & 0 & 0 & 0\end{pmatrix}. \quad (24)$$

Analogously for the orientation we have

$$\hat{\boldsymbol{t}}_s=\frac{\partial\hat{\boldsymbol{x}}_s}{\partial s}=\begin{pmatrix}3s^2\\ 2s\\ 1\\ 0\end{pmatrix}\cdot\boldsymbol{H}\cdot\begin{pmatrix}\boldsymbol{x}_j\\ \boldsymbol{x}_k\\ \boldsymbol{t}_j\\ \boldsymbol{t}_k\end{pmatrix}. \quad (25)$$

Note that the exact same formulae are used for interpolating curves between 3D primitives, but applied to 3 dimensions instead of 2.

The other modalities are interpolated by assuming that these change linearly with s between $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$:

Phase: The phase modality of the primitive interpolated for $s\in[0,1]$ is computed as by

$$\hat{\boldsymbol{\phi}}_s=\arctan\left(\frac{(1-s)\sin(\phi_j)+s\sin(\phi_k)}{(1-s)\cos(\phi_j)+s\cos(\phi_k)}\right). \quad (26)$$

Colour: The colour of the interpolated primitive is computed using the following equation:

$$\hat{\boldsymbol{c}}_s=(1-s)\boldsymbol{c}_j+s\boldsymbol{c}_k. \quad (27)$$

2D Primitive correction. We can then correct the *extracted* primitive $\boldsymbol{\pi}_i$ between $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$ with the *interpolated* primitive $\hat{\boldsymbol{\pi}}_s$. This is done for each modality m using a weighted mean between the two values. For position and colour information $m\in\{\boldsymbol{x},\boldsymbol{c}\}$, the corrected value $\overline{m}$ is computed by

$$\overline{m}_i=(1-\lambda)m_i+\lambda\hat{\boldsymbol{m}}_{s,j,k}, \quad (28)$$

where m_i is the extracted modality value, $\hat{\boldsymbol{m}}_{s,j,k}$ is the value interpolated at $\boldsymbol{x}_s$ between $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_k$, and λ is the correction rate.

For orientation and phase $m\in\{\theta,\phi\}$, we have:

$$\overline{m}_i=\arctan\left(\frac{(1-\lambda)\sin(m_i)+\lambda\sin(\hat{\boldsymbol{m}}_{s,j,k})}{(1-\lambda)\cos(m_i)+\lambda\cos(\hat{\boldsymbol{m}}_{s,j,k})}\right) \quad (29)$$

Note that in the case of $|\hat{\theta}-\theta|\leq\frac{\pi}{2}$, we need to operate a switch of the primitive's interpretation of the orientation as defined in Ref. [9] before correcting the orientation, colour and phase.

The correction (in Equations 28 and 29) is applied for N iterations, with a correction factor $\lambda=1/N$. This is evaluated on an artificial scene with precise 3D ground truth in the results section, and the results showed that a small number of iterations can already considerably improve accuracy.

3D primitive correction. In the 3D case, the primitives also suffer from the uncertainty that originates from the stereo matching and reconstruction processes. The 3D primitives' position in space is corrected to

$$\overline{\boldsymbol{X}}_i=(1-\lambda)\boldsymbol{X}_i+\lambda\hat{\boldsymbol{X}}_{s,j,k}, \quad (30)$$

and the orientation to

$$\overline{\boldsymbol{\Theta}}_i=\frac{(1-\lambda)\boldsymbol{\Theta}_i+\lambda\hat{\boldsymbol{\Theta}}_{s,j,k}}{\|(1-\lambda)\boldsymbol{\Theta}_i+\lambda\hat{\boldsymbol{\Theta}}_{s,j,k}\|}. \quad (31)$$

This correction is applied iteratively N times, with a correction factor $\lambda=1/N$. Also in this case, the results section shows that a small number of iteration suffice to improve accuracy.

Results

This section contains an evaluation of the different mechanisms presented above. In order to evaluate the performance of the different algorithms, we used stereo video sequences generated from a high resolution images of a urban scenes, with the associated depth ground truth provided with range scanner.

The range scanner provided us with a single high–resolution image with associated range information, and therefore each pixel of the image is given by

$$\mathbf{S}_{ij}=(X,Y,Z,r,g,b), \quad (32)$$

where (r,g,b) is the pixel's colour and (x,y,z) is the corresponding 3D point (according to the range scanner). For each image, we then define ten virtual pairs of stereo cameras with resolution 1024×1024, and used projective geometry to transform the original image pixels into the virtual cameras' images, then the colour of each pixel in the virtual images is linearly interpolated from the nearest 4 transformed points. The disparity between the two virtual stereo views is also linearly interpolated at all pixel positions — see Figure 9.

This offers realistic video sequences with an accurate 3D ground truth. Some images generated from three different range images are illustrated in Figure 10A, B and C; the dark blue areas (like the sky) correspond to where there was no range data available, and therefore the colour cannot be interpolated. No range data was available for sequence D, therefore we only have a qualitative evaluation on this sequence.

Stereo Evaluation

We first assessed the performance of the stereo matching scheme using each modal distance individually, plus the proposed multi–modal distance. We used the sequences with ground truth in Figure 10A, B, C to evaluate quantitatively the efficiency of each measure for stereo matching. We considered that a match was correct if its disparity error with the ground truth was smaller than

Figure 9. Illustration of how a sequence is generated from colour range images. The images show the first $t=1$ left and right images, the left disparity image, and the last left image ($t+\delta t=10$).
doi:10.1371/journal.pone.0010663.g009

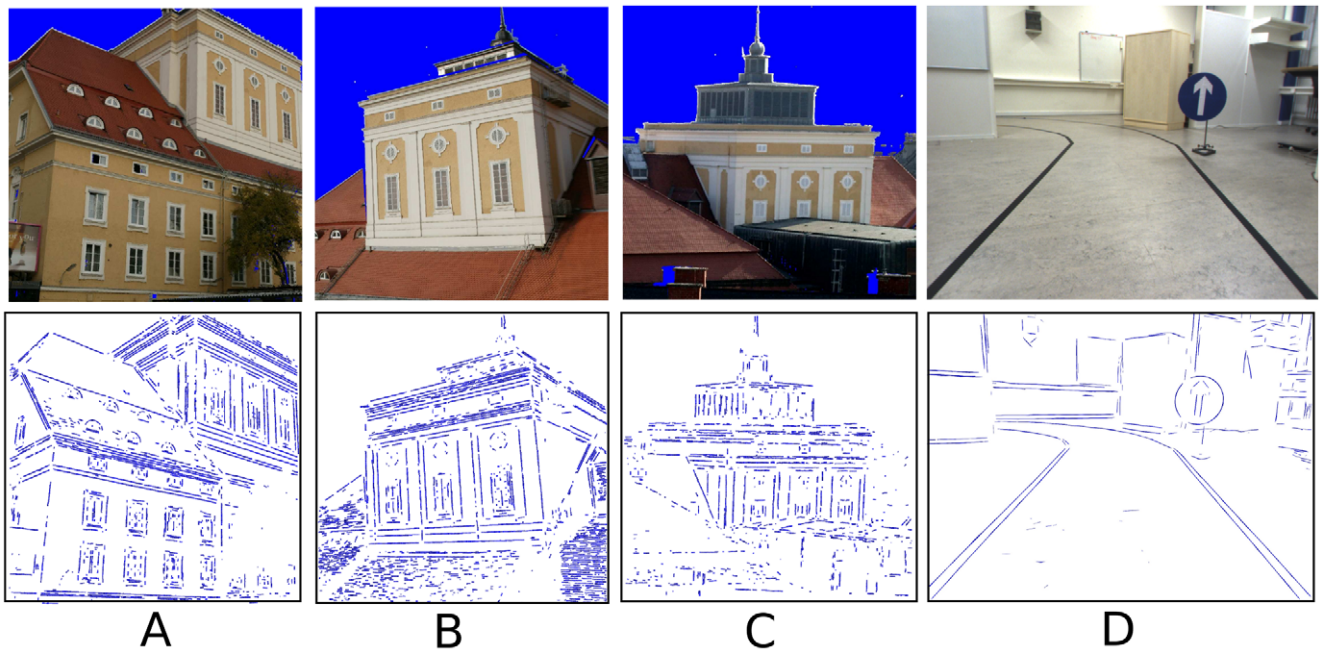

Figure 10. The four sequences on which we tested our approach. The top row shows one image from each sequence, and the bottom row shows the groups created.
doi:10.1371/journal.pone.0010663.g010

the 2D primitives' size — this ensures that no erroneous match is considered as correct.

Figure 11 shows the histogram distributions of the modal distances between primitive pairs satisfying the epipolar constraint — for all images in sequences A, B and C. All histograms show a separation between the distributions of correct (black) and false (white) correspondences. In the phase (Figure 11 top–right) and colour (Figure 11 bottom–left) histograms, the correct correspondences show a sharp peak at a modal distance of zero, whereas the false ones display an even distribution along all distances between [0,1]. In the orientation histogram (Figure 11 top–left), the large peak at zero distance for false correspondences is explainable by the presence of parallel structures in the image. Consequently, if one draws a horizontal line in the image, this line would cross parallel contours of very similar local orientation. The optical flow distribution shown in Figure 11 bottom–right has a peaked distribution centred at a distance of 0.1 for the correct correspondences, with a long tail until 0.6. The fact that the distribution peaks at 0.1 is explained by the projective difference in the optical flow between the two stereo images (the flow is likely to be similar, but not equal); this long tail is likely to be a consequence of the noisiness of optical flow data. The false correspondences also show a broad distribution around a modal distance of 0.3; the fact that the distribution is not centred at 0.5 is a consequence of statistical distributions of edges in natural images: horizontal and vertical edges are more likely, and therefore horizontal and vertical flow vectors are also more likely. In spite of this large overlap, optical flow distance is still better than chance for identifying correct stereo correspondences from erroneous ones — see ROC analysis in Fig. 12B: the optic flow curve is above the diagonal line that indicates chance performance in ROC curves. Figure 12A shows the multi–modal similarity histogram for correct and erroneous stereo matches. There is little overlap between the two distributions, showing that the multi–modal similarity is a good criterion for stereo matching.

In order to evaluate the performance of each distance measure for the task of identifying correct stereo matches from erroneous ones, we drew the Receiver Operating Characteristic (ROC) curves for each of them. If we consider a set of putative stereo correspondences, provided that we have a distance measure for all of them and that we know from the disparity ground truth which ones are correct, it is possible to compute the ratios of correct and erroneous pairs of

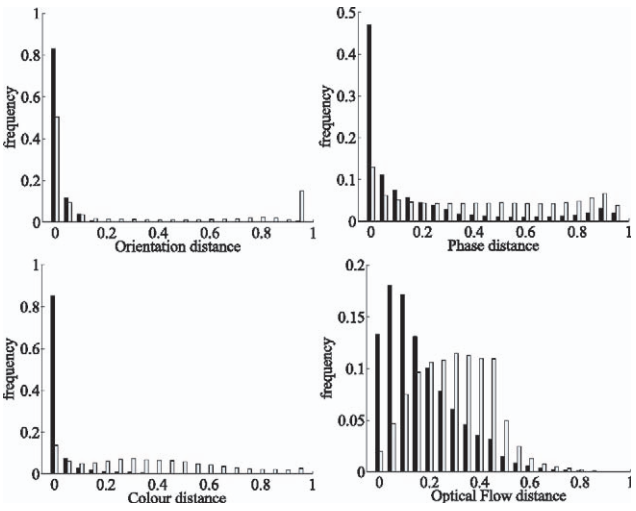

Figure 11. Histograms of the modal distances. Each plot shows the histograms of one modal distance (0 for identity and 1 for dissimilar items), for correct (black bars) and false (white bars) correspondences. The modal distances between putative stereo pairs are binned along the horizontal axis, and the vertical axis shows the frequency of occurrence of this value, between 0 and 1 (such that the cumulated heights of black and white bars are both 1). The histograms are computed across all three sequences in Figure 10 **A**, **B** and **C**.
doi:10.1371/journal.pone.0010663.g011

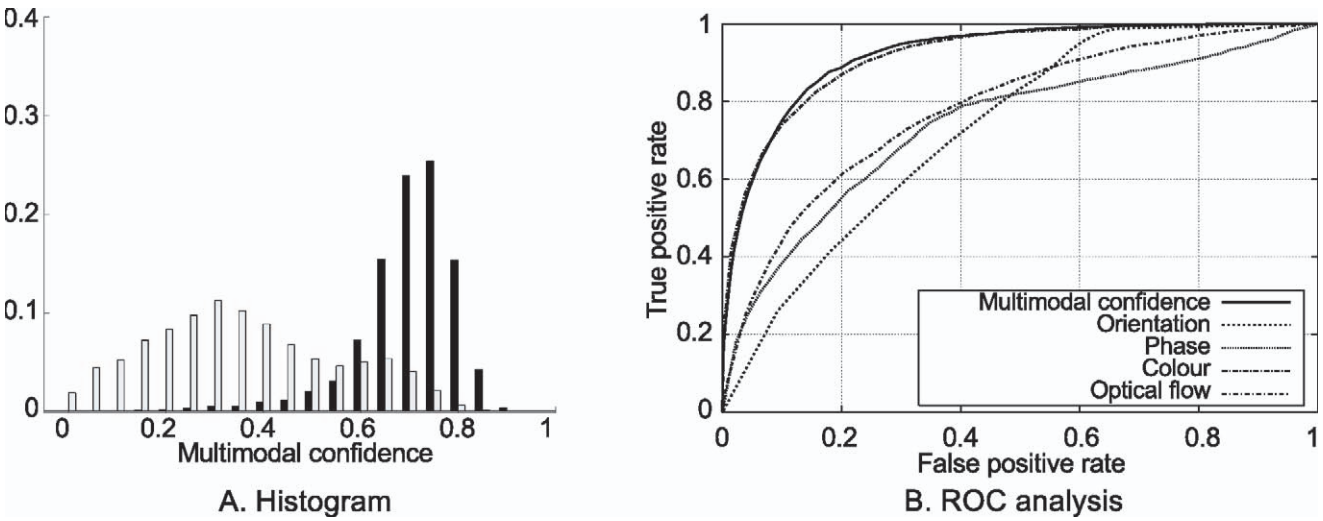

Figure 12. Evaluation of the multi–modal stereo. A Histogram of the multi–modal similarities between correct (black bars) and false (white bars) potential correspondences. **B** ROC curves for the different modalities. These results have been collected over 10 frames of the sequences Figure 10 **A**, **B** and **C**.
doi:10.1371/journal.pone.0010663.g012

primitives with a distance below threshold, respectively called *true* and *false positive rates*. A ROC curve records the true positive rates against the false positive rates obtained when considering one distance measure for a sample of threshold values ranging from 0 to 1. Therefore, a random measurement would generate a nearly diagonal ROC curve, whereas a measurement that is very significant for the task would have a large area below its ROC curve. In Figure 12B, such ROC curves show the performance of the stereo matching. Each of the curves shows the performance when using each modal similarity, or the multi–modal similarity proposed in Equation (15). 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 (if we neglect the V component and only keep the H and S), whereas 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). Moreover, those stereo pairs of images were interpolated from a single high–resolution image with range ground truth; thus, pixel colour is consistency is unaffected by illumination and therefore artificially high between left and right images. 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 objects. 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. These results show that (1) the similarity measures in all modalities are efficient (i.e., better than chance) indicators for stereo matching, (2) the multi–modal similarity yields a better classification.

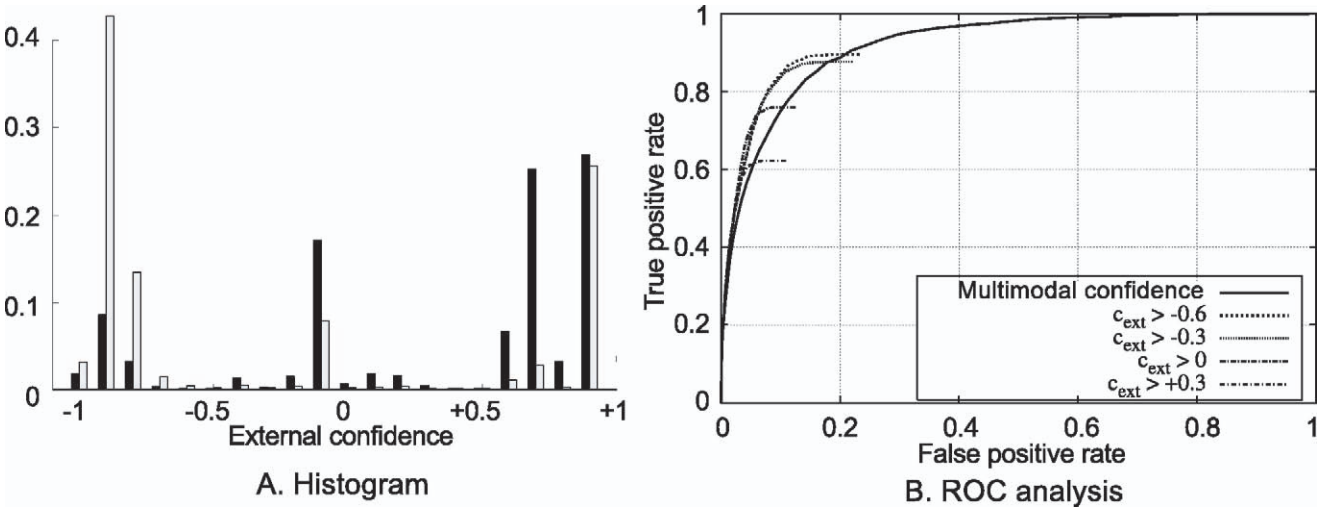

Figure 13. Evaluation of the external confidence. A Histogram of the external confidence rating for correct (black bars) and false (white bars) correspondences. **B** Each curve stands for a the application of a different threshold over the external confidence, prior to the ROC analysis. These curves represent the statistics over 10 frames of the three sequences with ground truth — see Figure 10 **A**, **B** and **C**.
doi:10.1371/journal.pone.0010663.g013

External Confidence Threshold

In a second set of experiments, we evaluated the effect of setting a minimal threshold on the external confidence. The external confidence threshold was always applied in conjunction with a sensible threshold on the multi–modal similarity of $\tau_m = 0.8$.

In Figure 13A, one can see that the correct (black) correspondences have mostly positive external confidences, while incorrect (white) ones have mainly negative values (large peak at -0.9). The small peak of correct correspondences for negative external confidence (near -0.9) is due to the few cases where most primitives on a contour have an erroneous correspondence, and therefore the few correct ones are strongly contradicted. The large values of erroneous correspondences with external confidences of 1 comes from repetitive structures in the image, that require more global considerations for disambiguation. 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 of zero implies the removal of all the isolated primitives (see constraint **C1**) as an isolated primitive has an external confidence of zero by definition.

Figure 13B shows ROC curves of the performance for varying thresholds on the multi–modal similarity. Each curve shows the performance for a different threshold (with threshold 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 these 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 — different thresholds can be chosen. The best overall improvement seems to be reached for a threshold of -0.6 over the external confidence (with a negligible difference in performance between -0.3 and -0.6). However, in the general case where a high reliability is required of the stereo matches, a small positive threshold of 0.1 is preferred (meaning discarding all primitives which are not part of a group) is preferred. Note that when a threshold is applied to the external confidence prior to the ROC analysis, the resulting curve does not reach the (1,1) point of the graph. This is normal as the threshold already removes some stereo hypotheses even before the multi–modal confidence is considered.

Table 1 summarises the performance of the stereo matching scheme, with and without external confidence threshold (because the external confidence is within $[-1,1]$, a threshold of -1 is the same as no threshold at all), on all three sequences with ground truth, showing a consistent improvement in all scenes, although the actual magnitude of the improvement varies. Sequence A, for example, contains a lot of repetitive, parallel structures which the external confidence cannot help disambiguating.

Table 1. Performance of the stereopsis with and without external confidence threshold.

sequence	τ_m	τ_e	correct c	false f	$\frac{c-f}{c+f}$
A	0.8	−1.0	3633	498	0.76
A	0.8	−0.1	3582	456	0.77
B	0.8	−1.0	2205	1178	0.30
B	0.8	−0.1	1915	447	0.62
C	0.8	−1.0	906	276	0.53
C	0.8	−0.1	804	167	0.66

$\tau_m \in [0,1]$ is the multi–modal similarity threshold for stereo matching; $\tau_e \in [-1,+1]$ is the external confidence threshold; c and f are the total number of true and false correspondences (respectively) selected by these thresholds.
doi:10.1371/journal.pone.0010663.t001

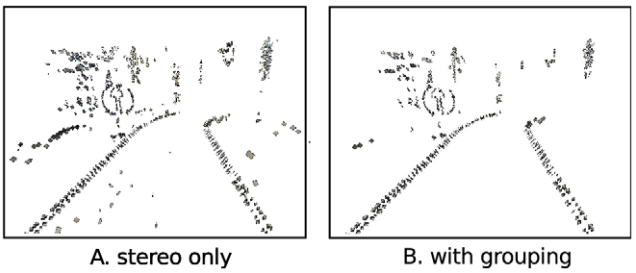

Figure 14. Qualitative example of the effect of the external confidence threshold. A primitives reconstructed from the sequence in Figure 10D, without threshold on external confidence ($\tau_m = 0.8$, $\tau_e = -1.0$). **B** primitives reconstructed from the same sequence with a threshold on external confidence ($\tau_m = 0.8$, $\tau_e = 0.1$).
doi:10.1371/journal.pone.0010663.g014

Figure 14 illustrates the effect qualitatively for the video sequence from Figure 10D. Figure 14a) shows the 3D primitives reconstructed with a threshold on external confidence of $\tau_e = -0.1$. When comparing Figures 14A and 14B we can see that a large number of outliers has been discarded from the reconstructed 3D primitives, leading to a cleaner description of the scene.

Interpolation

We evaluated the performance of the interpolation scheme, on two simple artificial sequences illustrated in Figure 15. In the case of 3D–interpolation we also evaluated the interpolation effect on the reconstructed 3D representation qualitatively. The interpolation scheme was applied for $N = 10$ iterations, with a correction factor of $\lambda = 0.1$.

2D interpolation Results. The results for localisation, orientation and phase over 10 iterations of the correction process are shown in Figure 16, for the triangle (full line) and the circle (dashed line) scenarios. The horizontal axis shows the number of iterations of the correction process and the vertical axis the mean error of the 2D primitives. Note that the error is measured in pixels for the localisation and in radians for the orientation and the phase.

This sub–pixel accuracy is naturally lower for the circle scene, which is due to the contour's curvature. As primitives are local line descriptors, they can describe curved contours but they assume

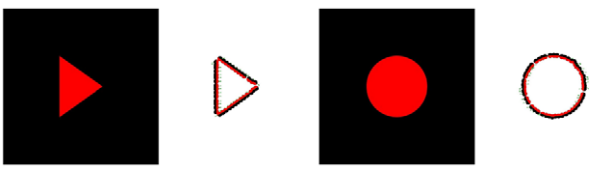

Figure 15. Illustration of the primitives extracted from two simple artificial sequences, featuring a triangle (left) and a circle (right). In both scenarios, the object (triangle or circle) is facing the cameras, at a depth of 100 units, the object has a radius of 10 units, and the baseline between the two cameras is 10 units. Both images shown here are from the last camera.
doi:10.1371/journal.pone.0010663.g015

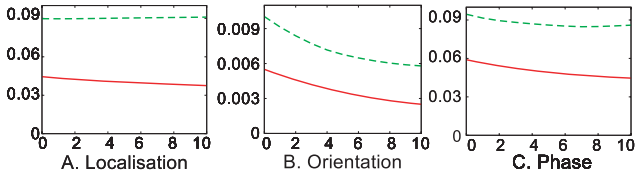

Figure 16. Correction of the 2D primitives using interpolation. Accuracy of the 2D primitives' localisation (**A**), orientation (**B**) and phase (**C**) after several iterations of the correction process, for the triangle (full line) and circle (dashed line) scenarios. The horizontal axis shows the number of iterations of the correction process and the vertical axis shows the error for **A** in pixels, and for **B** and **C** in radians.
doi:10.1371/journal.pone.0010663.g016

low local curvature. Hence, as the sub–pixel accuracy is assuming this linear model, it is performing better with purely linear structures. Nonetheless, note that the accuracy is extremely high in both cases: less than one tenth of a pixel for the localisation and and less than one hundredth of a radian for the orientation — i.e., less than 0.6 degrees.

Moreover, we note that interpolation leads to mixed results depending on the modality: we see a distinct improvement of the localisation for the triangle scene, but not for the circle scene. This is likely to be due to the use of Hermite interpolation, in two respects: first, Hermite interpolation makes use of the tangents' orientation in addition to their position; hence, the interpolated curve is sensitive to errors in orientation. Second, even if the Hermite polynomials are an efficient model for describing general curves, they do not allow a perfect interpolation of an arc; thus, interpolation at high curvature locations lead to a loss in precision. Nonetheless, the accuracy of the interpolated primitive itself is always better than the original (reconstructed by stereo).

Concerning orientation, we see a clear improvement of $\sim$0.003 radians for both objects ($\sim$50% and $\sim$30% for the triangle and circle). Phase shows a clear (although smaller) improvement in both cases; the triangle scenario sees an improvement of $\sim$0.015 ($\sim$25%), whereas the circle scenario sees an improvement of $\sim$0.01 ($\sim$11%). The effect of phase correction is illustrated in Figure 17. This figure shows a detail of the primitives extracted on the circle scene; the phase is illustrated on the primitives by the green arrow, which orientation indicates the phase. In this case, horizontal indicates a full contrast edge structure, and vertical a full contrast line. Figure 17C and D show the phase before and after correction, where the dotted lines show the mean phase across the whole circle. Before correction, the phase of the central primitive differs significantly from the correct one, and it is closer to the dotted line after correction.

3D primitives interpolation. This scheme was evaluated on the same triangle sequence as above (shown in Figure 15) and

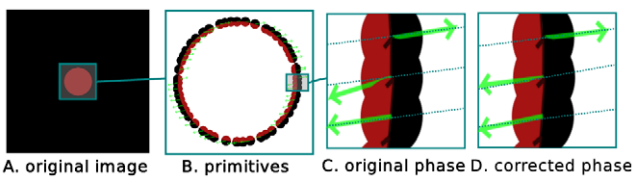

Figure 17. Illustration of the effect of phase correction in 2D. A the original image; **B** the extracted primitives; **C** detail of the primitives, the green arrows show the extracted phase, the dotted lines show the mean phase over the whole circle; **D** detail of the primitives after correction: the central primitive's phase is now closer to the dotted line.
doi:10.1371/journal.pone.0010663.g017

Table 2. Effect of the correction process on the localisation and orientation in space of the primitives reconstructed from the triangle scenario.

	localisation error		orientation error	
	mean	variance	mean	variance
before	0.03524	0.00392	0.01712	0.00082
after 10 iterations	0.02426	0.00221	0.01434	0.00056

doi:10.1371/journal.pone.0010663.t002

resulted in a reduction of the localisation error by $\sim$30%; the orientation error was reduced by $\sim$16% (see Table 2). When applying the same scheme to the circle scenario, the localisation error was reduced by $\sim$20%; orientation error was reduced also by $\sim$20% (see Table 3 and Figure 18). Figure 19 shows the effect of this smoothing on selected details in an indoor scene.

Discussion

In this paper, we presented several local operations on the visual primitives presented in Ref. [9], which produce a robust representation of visual scenes, some of them making use of the (still locally constrained) context.

First, we presented a simple algorithm to group primitives into contours. Contours were defined implicitly in terms of the pairwise relations between proximate 2D primitives. Note that an explicit description of the groups could easily be extracted from such an implicit definition using a variety of techniques, including: normalised [52] or average cuts [53], affinity normalisation [15], dynamic programming [54], probabilistic chaining [55], etc.

Second, we proposed to use the multi–modal similarity between 2D primitives to perform stereo matching between pairs of images. The stereo algorithm we used is purely local and therefore does not make use of global constraints (e.g., ordering constraint [56], figural continuity [57], etc.), or optimisation (e.g., dynamic programming [58], graph operations like maximal clique [59], etc.). Such global optimisations generally allow to improve significantly the performance of local stereo matching schemes, and therefore could be applied to this system to further improve the quality of stereo matching.

Third, we proposed a scheme integrating contextual information combining perceptual grouping and stereopsis to improve the reliability of the latter. The external confidence defined here is comparable to averaging over a local neighbourhood of a disparity gradient constraint along contours [60]. Also, in a similar way, Ohta and Kanade [56] proposed to apply inter–scanline consistency rules in addition to a more classical intra–scanline ordering constraint. Departing from those pixel–based constraints,

Table 3. Effect of the correction process on the localisation and orientation in space of the primitives reconstructed from the circle scenario.

	localisation error		orientation error	
	mean	variance	mean	variance
before	0.08653	0.01188	0.02476	0.00071
after 10 iterations	0.06868	0.00882	0.01955	0.00046

doi:10.1371/journal.pone.0010663.t003

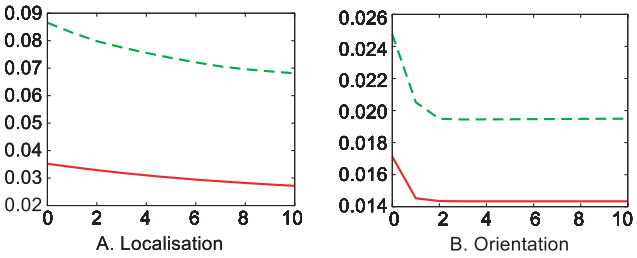

Figure 18. Correction of 3D primitives. Error of the **A** localisation and **B** orientation of the reconstructed 3D primitives after several iterations of the correction process. Solid lines shows the errors for the triangle scenario and dashed line for the circle scenario. The horizontal axis shows the number of iterations of the correction process and the vertical axis shows the error in **A** units (in the 3D space, arbitrary in an artificial scenario) and **B** radians.
doi:10.1371/journal.pone.0010663.g018

the definition of the Basic Stereo Consistency Event (BSCE) allows to specify semantically which neighbours have positive and negative contributions to the confidence. It was shown that it could improve significantly the reliability of stereo matching.

Moreover, we showed that the same grouping relation can be used to interpolate contours between pairs of linked primitives. This was then used to correct primitives with the contour as interpolated from its neighbours. In 2D, we obtained a reduction by more than 30% of the orientation error, and more than 10% for the phase. When interpolating 3D primitives, we additionally found that the localisation error was reduced by more than 20%, and the orientation error by more than 15%. Therefore, this interpolation step proved to be a robust manner to improve the representation accuracy, both in 2D and 3D. Because the scheme is local, there is no *a priori* assumption that the whole contours comply with a certain mathematical description: we only assume that the contour is smooth between two proximate primitives, and model this using Hermite interpolation.

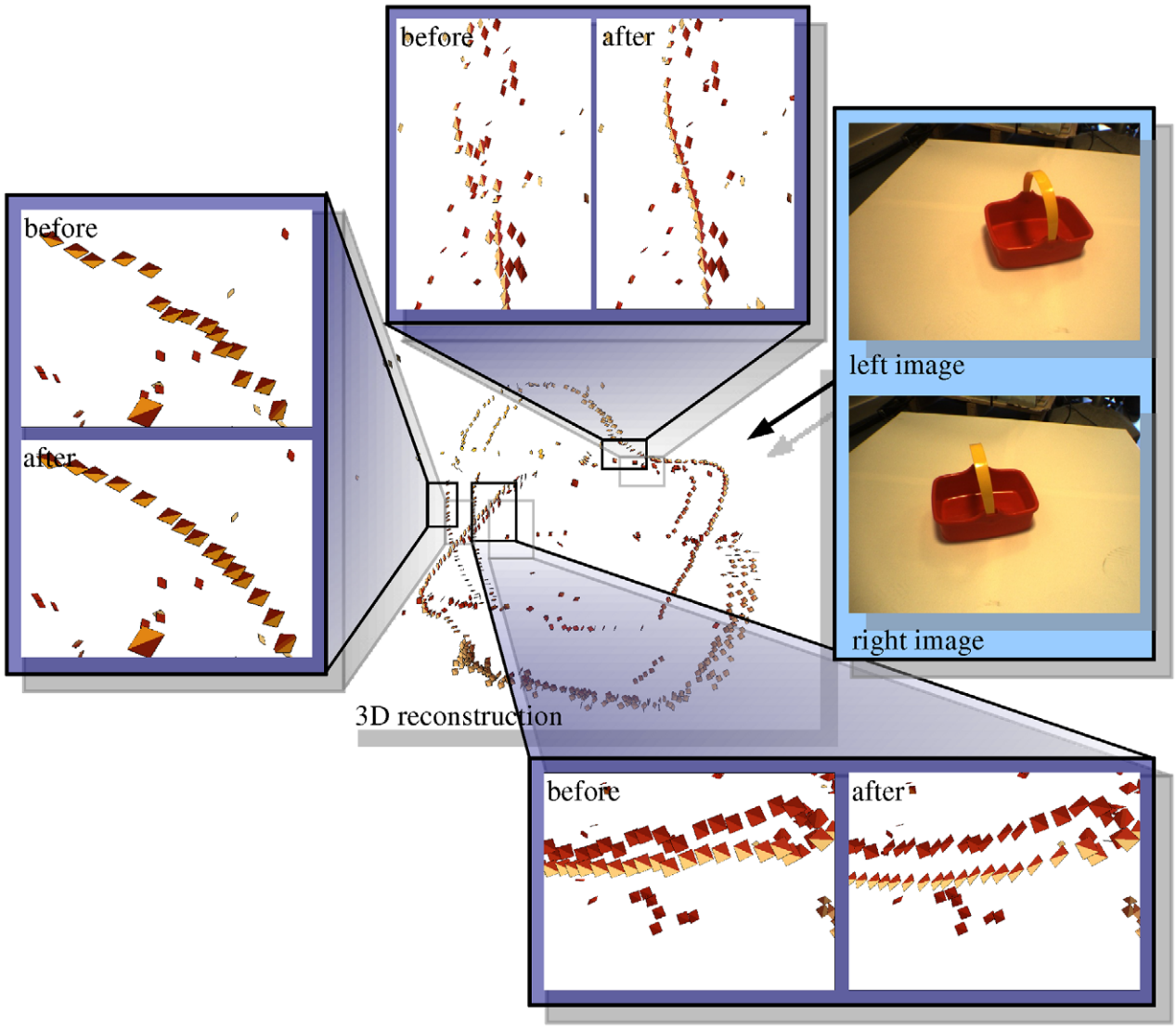

Figure 19. Illustration of the effect of the correction of 3D primitives using interpolation. The figure shows the reconstructed primitives before and after 10 rounds of correction, for details of an object.
doi:10.1371/journal.pone.0010663.g019

Finally, we showed that using such mutual feedback between mid–level, local processes allow to disambiguate them without need for additional contextual knowledge. Thereby, we provide a reliable 3D representation of the shapes in the scene that can then be used for higher level visual operations, where contextual knowledge may be available. This framework was used successfully to address a variety of robot vision tasks: e.g., grasping [13], ego–motion estimation [61], and learning of objects' shapes [12].

Acknowledgments

We thank the company RIEGL-UK Ltd. for the images with known ground truth used for sequence A, B and C.

Author Contributions

Conceived and designed the experiments: NP FW NK. Performed the experiments: NP. Analyzed the data: NP. Contributed reagents/materials/analysis tools: NP NK. Wrote the paper: NP FW NK.

References

1. Oram M, Perrett D (1994) Modeling visual recognition from neurobiological constraints. Neural Networks 7: 945–972.
2. Aloimonos Y, Shulman D (1989) Integration of Visual Modules — An Extension of the Marr Paradigm. Academic Press, London.
3. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27: 1615–1630.
4. Se S, Lowe D, Little J (2001) Vision-based mobile robot localization and mapping using scale-invariant features. In: IEEE International Conference on Robotics and Automation. volume 2. pp 2051–2058.
5. Lowe D (1999) Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision (ICCV'99). pp 1150–1157.
6. Mohan R, Nevatia R (1992) Perceptual organization for scene segmentation and description. IEEE Transactions on Pattern Analysis and Machine Intelligence 14: 616–635.
7. Chung R, Nevatia R (1995) Use of monucular groupings and occlusion analysis in a hierarchical stereo system. Computer Vision and Image Understanding 62: 245–268.
8. Geman S, Bienenstock E, Doursat R (1995) Neural networks and the bias/variance dilemma. Neural Computation 4: 1–58.
9. Krüger N, Lappe M, Wörgötter F (2004) Biologically motivated multi-modal processing of visual primitives. Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour, AISB Journal 1: 417–427.
10. Baseski E, Pugeault N, Kalkan S, Kraft D, Wörgötter F, et al. (2007) A scene representation based on multi-modal 2d and 3d features. In: ICCV Workshop on 3D Representation for Recognition 3dRR-07.
11. König P, Krüger N (2006) Perspectives: Symbols as self-emergent entities in an optimization process of feature extraction and predictions. Biological Cybernetics 94: 325–334.
12. Kraft D, Pugeault N, Başeski E, Popović M, Kragic D, et al. (2009) Birth of the object: Detection of objectness and extraction of object shape through object action complexes. Special Issue on "Cognitive Humanoid Robots" of the International Journal of Humanoid Robotics 5: 247–265.
13. Popović M, Kraft D, Bodenhagen L, Başeski E, Pugeault N, et al. (accepted) A strategy for grasping unknown objects based on co–planarity and colour information. Robotic and Autonomous Systems.
14. Parent P, Zucker S (1989) Trace inference, curvature consistency, and curve detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 11: 823–839.
15. Perona P, Freeman W (1998) A Factorization Approach to Grouping. In: Proceedings of the 5th European Conference on Computer Vision (ECCV'98), LNCS 1406. volume 1. pp 655–670.
16. Elder J, Goldberg R (2002) Ecological statistics of Gestalt laws for the perceptual organization of contours. Journal of Vision 2: 324–353.
17. Krüger N, Felsberg M (2004) An explicit and compact coding of geometric and structural information applied to stereo matching. Pattern Recognition Letters 25: 849–863.
18. Pugeault N, Krüger N (2003) Multi–modal matching applied to stereo. Proceedings of the BMVC 2003: 271–280.
19. Schmid C, Mohr R, Baukhage C (2000) Evaluation of Interest Point Detectors. International Journal of Computer Vision 37: 151–172.
20. Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference. pp 147–151.
21. Lowe D (2004) Distinctive Image Features from Scale–Invariant Keypoints. International Journal of Computer Vision 60: 91–110.
22. Kovesi P (1999) Image features from phase congruency. Videre: Journal of Computer Vision Research 1: 1–26.
23. Rodrigues J, du Buf J (2006) Multi–scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. Biosystems 86: 75–90.
24. Rodrigues J, du Buf J (2009) Multi–scale lines and edges in V1 and beyond: brightness, object categorization and recognition, and consciousness. Biosystems 95: 206–226.
25. Baumberg A (2000) Reliable feature matching across widely separated views. In: Proceedings of the International Conference on Pattern Recognition. pp 774–781.
26. Koenderink J, van Doorn A (1987) Representation of Local Geometry in the Visual System. Biological Cybernetics 55: 367–375.
27. Schaffalitzky F, Zisserman A (2002) Multi–view matching for unordered image sets, or "how do I organize my holiday snaps?". Lecture Notes in Computer Science 2350: 414–431.
28. Freeman W, Adelson E (1991) The design and use of steerable filters. IEEE transactions on Pattern Analysis and Machine Intelligence 13: 891–906.
29. van Gool L, Moons T, Ungureanu D (1996) Affine / Photometric Invariants for Planar Intensity Patterns. Lecture Notes In Computer Science 1064: 642–651.
30. Elder J (1999) Are edges incomplete? International Journal of Computer Vision 34: 97–122.
31. Felsberg M, Sommer G (2001) The monogenic signal. IEEE Transactions on Signal Processing 49: 3136–3144.
32. Felsber M, Kalkan S, Krüger N (2009) Continuous dimensionality characterization of image structure. Image and Vision Computing 27: 628–636.
33. Nagel HH (1987) On the estimation of optic flow: Relations between different approaches and some new results. Artificial Intelligence 33: 299–324.
34. Koffka K (1935) Principles of Gestalt Psychology. Lund Humphries, London.
35. Köhler K (1947) Gestalt Psychology: An introduction to new concepts in psychology. New York: Liveright.
36. Wertheimer M, ed (1935) Laws of Organsation in Perceptual Forms. Harcourt & Brace & Javanowitch, London.
37. Field D, Hayes A, Hess R (1993) Contour integration by the human visual system: Evidence for a local "association field". Vision Research 33: 173–193.
38. Brunswik E, Kamiya J (1953) Ecological cue–validity of 'proximity' and of other Gestalt factors. American Journal of Psychology 66: 20–32.
39. Elder J, Goldberg R (1998) Inferential reliability of contour grouping cues in natural images. Perception Supplement 27.
40. Geisler W, Perry J, Super B, Gallogly D (2001) Edge Co–occurrence in Natural Images Predicts Contour Grouping Performance. Vision Research 41: 711–724.
41. Krüger N (1998) Collinearity and parallelism are statistically significant second order relations of complex cell responses. Neural Processing Letters 8: 117–129.
42. Faugeras O (1993) Three–Dimensional Computer Vision MIT Press.
43. Hartley R, Zisserman A (2000) Multiple View Geometry in Computer Vision Cambridge University Press.
44. Brown M, Burschka D, Hager G (2003) Advances in Computational Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 25: 993–1008.
45. Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two–frame stereo correspondence algorithms. International Journal of Computer Vision 47: 7–42.
46. Pugeault N (2008) Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation. Ph.D. thesis, University of Göttingen.
47. Pugeault N, Wörgötter F, Krüger N (2006) Multi-modal scene reconstruction using perceptual grouping constraints. In: Proc. IEEE Workshop on Perceptual Organization in Computer Vision (in conjunction with CVPR'06). pp 195–213.
48. Pugeault N, Kalkan S, Başeski E, Wörgötter F, Krüger N (2008) Reconstruction uncertainty and 3D relations. In: Proceedings of Int. Conf. on Computer Vision Theory and Applications (VISAPP'08). volume 2. pp 186–193.
49. Verri A, Torre V (1986) Absolute depth estimate in stereopsis. Journal of the Optical Society of America 3.
50. Wolff L (1989) Accurate measurements of orientation from stereo using line correspondence. In: Proceedings of the IEEE Computer Vision and Pattern Recognition conference (CVPR'98). pp 410–415.
51. Wikipedia (2007) Cubic Hermite Spline. URL http://en.wikipedia.org/wiki/Cubic_Hermite_spline.
52. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22: 888–905.
53. Sarkar S, Soundararajan P (2000) Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. IEEE Transactions on Pattern Analysis and Machine Intelligence 22: 504–525.
54. Sha'ashua A, Ullman S (1990) Grouping contours by iterated pairing network. In: Neural Information Processing Systems (NIPS). volume 3. pp 335–341.
55. Crevier D (1999) A probabilistic method for extracting chains of collinear segments. Computer Vision and Image Understanding 76: 36–53.
56. Ohta Y, Kanade T (1985) Stereo by intra– and inter–scanline search using dynamic programming. IEEE transactions on Pattern Analysis and Machine Intelligence 7.
57. Mayhew J, Frisby J (1981) Psychophysical and computational studies towards a theory of human stereopsis. Artificial Intelligence 17: 349–385.

58. Lee SH, Leou JJ (1994) A dynamic programming approach to line segment matching in stereo vision. Pattern Recognition 27: 961–986.
59. Horaud R, Skordas T (1989) Stereo correspondences through feature grouping and maximal cliques. IEEE transactions on Pattern Analysis and Machine Intelligence 11.
60. Kim N, Bovik A (1988) A contour–based stereo matching algorithm using disparity continuity. Pattern Recognition 21: 505–514.
61. Pugeault N, Wörgötter F, Krüger N (2006) Rigid body motion in an early cognitive vision framework. In: Proceedings of the IEEE SMC UK&RI Conference on Advances in Cybernetic Systems (AICS'06). pp 217–223.