Brain Inspired Cognitive Systems
August 29-September 1, 2004
University of Stirling, Scotland, UK

Grouping Over Stereo for Visual Cues Disambiguation

Nicolas Pugeault[1], Florentin Wörgötter[1] and Norbert Krüger[4]

[1] Psychology, University of Stirling, Scotland
nicolas,worgott@cn.stir.ac.uk
[2] Computer Sci., Aalborg University Esbjerg, Denmark
nk@cs.aue.auc.dk

ABSTRACT

In stereo–vision, the goal is to reconstruct the three–dimensional structure of the scene observed from two camera inputs. The core problems are the matching of features into both camera frames, and the interpretation of image features in terms of the 3D scene. In this paper, we use a rating scheme of the potential correspondences, based on the multi–modal intrinsic similarity of the features. We propose here to join approach of stereo feature matching and feature grouping processes, into one intertwined spatial and stereo integrated process. We show here that those two apparently separated processes are based on the same assumptions. This joint approach allow to improve reliability and performance of both processes, and to solve some of their inherent ambiguities

1 Introduction

We describe a scene representation based on Primitives maps — see figure 1. Image Primitives are the combined product of local filters, sampled at points of interest : area in the image likely to contain manifestations of 3D features, concretely edges and corners. Those Image Primitives are local, sparse, and multi–modal. They code local estimations of established visual sub–structures such as orientation, contrast transition, color and optic flow in a condensed way (KLW03). Consequently, our stereo is processed on this Image Primitive level, matching the multi–modal information, and only processing the matching at those interesting points. We can show that it is the use of all modalities that results in the best performance — see (KF04) and (PK03).

The result of our stereo is more than a standard depth map that stores a depth value for each pixel. We yield a map that consists of descriptors with higher semantic value than a standard depth map. We call these descriptors Spatial Primitives. In addition to the 3D position, Spatial Primitives carry an additional geometric attributes in terms of a 3D orientation. Furthermore, the condensed information about the structural properties of the underlying 2D attributes colour and contrast transition are associated to the Spatial Primitives. Therefore, the Spatial Primitives carry information about aspects that are referred to as geometric and appearance based. We think that the debate about these different characteristics of visual information should not focus on their different importance but more on their different role in vision — as discussed, e.g., in (MLP^{+}96). Since both aspects are relevant it is essential to code both of them efficiently as done in the Primitives.

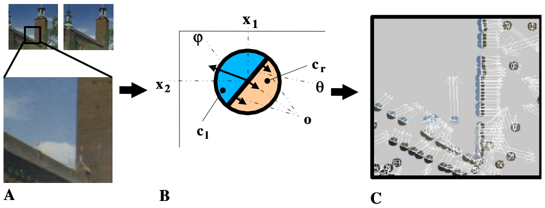

Fig. 1. Our primitive extraction applied to a sequence of two frames. The two frames are shown at the left. In the middle a schematic representation of a basic feature vector. Position is coded by (x, y), orientation by θ, phase by ϕ, and colour by (c_l, c_r), the colour on both sides of the edge. The right image shows the primitives and their modalities drawn at the position where they have been extracted.

As a consequence of their rich semantic, the Primitives allow for rich predictions in their spatial and spatial–temporal context — see respectively (KW02) and (KJP02). Here, we make explicit use of this potential by extrapolating Primitives including all their geometrical and structural attributes from consistent collinear tuplets and apply that to extract more reliable scene representations.

Ambiguity appears in many forms in visual processing (see, e.g., (AS89; KW04)) and may be caused by spurious noise, illumination, accidental background constellations or under scale features. In all those cases the resulting Primitive, if fundamentally correct in terms of signal, is improper for the study and analysis of the scene at the current scale. It proves impossible to indentify these ambiguities using only local information.

An additional sort of ambiguity is added through stereo. For stereo being processed on a local level all incorrect correspondences will produce falses 3D features in the reconstruction. The resulting map of 3D pseudo–features can be extremely confusing and little can be done at this stage to remove such an amount of noise.

We propose here to improve scene representations by applying new constraints to the Image Primitives. Constraints we believe are imposed by natural scene statistics, physiology as well as the geometric properties of stereo–vision. Our primitive being local estimation of edges, we consider that, assuming an adequate scale, any important structure of the scene will be represented by more than one Image Primitive in the image representation. Consequently, we are interested in groups of features as manifestations of scene structure, more than Primitives themselves, which are local sampling of those structures. Reversely, an organized constellation of primitives can be assumed to be the manifestation of a feature of the scene — other cases are typical of visual illusions.

Thus, we also want to apply this constraint to stereo, according to the following rule:

> **Stereo Consistency Constraint:** If Primitives form a group in the left image, and if those primitives have stereo correspondences, then those correspondences should form a group in the right image.

Our processing is essential hierarchical, however with a hierarchy that makes essential use of feedback loops to modify and correct decisions on earlier stages of processing by making use of structural knowledge gained at higher stages.

More specifically, we transform our data from:

Pixel Level: raw pixel information, noisy and relative to the point of view and optic of the camera
Image Primitives: to Image Primitives, containing meaningful information about the structure of the image
Potential Spatial Primitives: to a probabilistic Spatial Primitive map making use of multi–modal stereo
Stereo Primitive Constellations: holding groups of Image Primitives consistent in space and over stereo

There are recurrent processes between most of these stages. For example, the grouping process starts in the 2D domain, is then used to extract Spatial Primitives and then reflects back to extract more precise Image Primitives. We will show that this hierarchical and recurrent processing scheme improves significantly the quality of the scene representation. As a result it is not a pixel to pixel depth map we obtain, but a consistent reconstruction of meaningful features in 3D, extrapolated from the Image Primitive maps.

2 stereopsis

We want to use our image representation to compute stereopsis. If we consider that we have a pair of calibrated camera, we can apply our preprocessing to both video feed and apply stereo matching to our image representations. Through those stereo Primitives we can then reconstruct 3D elements, or Spatial Primitives. Those Spatial Primitives are also multi–modal and can be seen as local estimation of scene features.

2.1 The Stereo Algorithm

For each Image Primitive in the left image, the *epipolar line* is computed. Then all Image Primitives of the right image are considered. Their position is estimated in tangential and normal components to the epipolar line. The normal component being the distance to the line and the tangential the distance from the left Image Primitive to the projection. Effectively the normal component is the imprecision in the matching, due to sparseness, and the tangential is the disparity. All Image Primitives in the right frame close enough to the epipolar line are considered as *potential* matches for the left Image Primitive. Excluding cases of occlusion, unsolvable through local approaches, the main problem of stereo matching is to find the correct correspondences out of those hypothesis. First the sparseness of our image representation allows to focus on important and solvable areas — no accurate stereo matching can be processed on intrinsicly one dimensional areas.

3 Formalization and Use of Image Primitive Consistency

We want to define group of locally consistent Image Primitives in the image. We are interested in Image Primitives outlining major structures of the scenery, and subsequently of the images processed. We assume that any structure of the scene having a projective manifestation in the image, has a representation involving a set of consistent Image Primitives — in the following called group. From this assumption follows naturally that Image Primitives showing inconsistency with their neighbourhood might be considered as ambigious information likely to be caused by erroneous feature extraction. This confidence is evaluated through the following criteria: a *proximity* constraint: the primitive should distant by less than 8 times the size of the primitive patch — for psychophysical justification see (FHH92). Secondly a *collinearity* constraint: the second primitive should form an angle of less than 45 degrees with the orientation of the first one. Finally a *modality continuity* constraint: the visual modalities of the two primitives should be similar.

4 Primitive Linking

4.1 Links: Image Primitive Consistency Units

Those *Links* represent the overall compatibility of two Image Primitives, in a multi–modal sense.

We want to define group of locally consistent Image Primitives in the image. We are interested in Image Primitives outlining major structures of the scenery, and subsequently of the images processed. We assume that any structure of the scene having a projective manifestation in the image, has a representation involving a set of consistent Image Primitives — in the following called group. From this assumption follows naturally that Image Primitives showing inconsistency with their neighbourhood might be considered as ambigious information likely to be caused by erroneous feature extraction. Now, we want to define the meaning of this *consistency* in the multi–modal space of the features.

In this work, we consider Image Primitives defining local oriented structures — *i. e.* lines and step edges. Therefore, we are looking for Constellations defining global contours. Consistency between two Image Primitives is defined by two criterions: Collinearity and Modality Consistency — using the modalities colour and contrast transition. Inconsistency according to these two criterions indicates that the two Image Primitives are either expressions of independent structures or caused by the erroneous feature extraction process. In the following formulas we will consider a pair of Image Primitives $\mathbf{e}_1, \mathbf{e}_2$ such as $\mathbf{e}_2 \in N(\mathbf{e}_1)$, N being a large enough neighbourhood. We want to define relationships between $\mathbf{e}_1$ and $\mathbf{e}_2$ defining possible structures for $\mathbf{e}_1$ and we code them as links $l(\mathbf{e}_1, \mathbf{e}_2)$ between them. We associate a confidence $c[l(\mathbf{e}_1, \mathbf{e}_2)]$ to a Link which is an estimate of the probability for the two primitives to be part of the same structure.

This confidence is evaluated through the following criteria:

- a *proximity* constraint: the primitive should distant by less than 8 times the size of the primitive patch — for psychophysical justification see (FHH92).
- a *collinearity* constraint: the second primitive should form an angle of less than 45 degrees with the orientation of the first one.
- a *modality continuity* constraint: the modalities of the two primitives should be similar.

4.2 Basic Stereo Consistency Event

We then define that the *minimal* stereo event involving a primitive neighbourhood, is: Given two Image Primitives $\mathbf{e}_1^L$ and $\mathbf{e}_2^L$ in the left frame such as a link $l(\mathbf{e}_1^L, \mathbf{e}_2^L)$ can be defined between them, if we consider the hypothesis that $s_i(\mathbf{e}_1^L)$ is the correct stereo–correspondence for $\mathbf{e}_1^L$ in the right image:

> **if** exists a link $l(s_i(\mathbf{e}_1^L), s(\mathbf{e}_2^L))$ between this stereo–correspondence and the public stereo–correspondence $s(\mathbf{e}_2^L)$ of the second primitive $\mathbf{e}_2^L$
> **then** the hypothesis $s(\mathbf{e}_1^L)$ is confirmed — and conversely if no corresponding link exist in the right image this hypothesis is then contradicted.

We call this trial the *Basic Stereo Consistency Event* (BSCE).

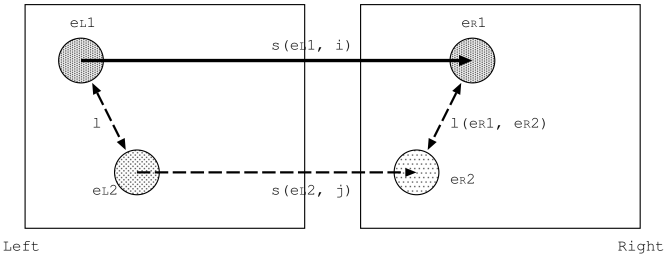

Fig. 2. The BSCE criterion: Given a stereo correspondence $s_i(\mathbf{e}_1)$, the BSCE can be calculated for a primitive $\mathbf{e}_2$ in the neighbourhood, depending on $l(\mathbf{e}_1, \mathbf{e}_2)$, $s_j(\mathbf{e}_2)$, and $l'(s_i(\mathbf{e}_1), s_)(\mathbf{e}_2 j)$. The bold line represent the event we want to confirm, and the dashed lines the external events which, in conjonction, confirms it.

4.3 Neighbourhood Consistency Confidence

This formula gives us how a Image Primitive stereo correspondence is consistent with our beliefs on another Image Primitive stereo properties. We now want to estimate how this correspondence is consistent with the *whole neighbourhood* of the Image Primitive. Now if we consider a primitive $\mathbf{e}_1^L$ and an associated stereo–correspondence $s_i(\mathbf{e}_1^L)$, we can integrate this BSCE confidence over the neighbourhood of the primitive ($N_{\mathbf{e}_1^L}$). We call this confidence the *external confidence* in the stereo–correspondence:

$$c_{ext}[s_i(\mathbf{e}_1^L)] = \frac{1}{|N_{\mathbf{e}_1^L}|} \sum_{\mathbf{e}_k^L \in N_{\mathbf{e}_1^L}} c[BSCE_i(\mathbf{e}_1^L, \mathbf{e}_k^L)] \qquad (1)$$

This gives us a confidence on how consistent is a stereo–correspondence with the stereo of the primitive neighbourhood.

5 Image Primitives Constellations and Interpolation

We then want to be able not only to discard incorrect hypothesis, but also to correct or to generate stereo hypothesis from a primitive neighbourhood. In order to achieve that we need an overdetermined system. Here we will focus on Triplets, which are fondamentally constituted of one central Image Primitive and two Links of this Image Primitive. For $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ Image Primitives of the image, if the links $l(\mathbf{e}_1, \mathbf{e}_2)$ and $l(\mathbf{e}_1, \mathbf{e}_3)$ exist, then we can define a triplet $T(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$

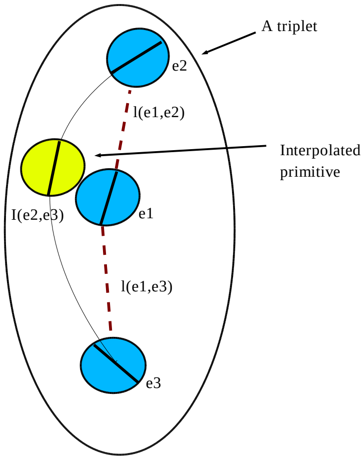

Fig. 3. The Triplet interpolation.

5.1 Interpolation

Given two consistent Image Primitives $\mathbf{e}_2, \mathbf{e}_3$, consistent in the sense that they are manifestation of the same scene feature, we can interpolate the feature at anypoint between them, using the Hermite interpolation scheme. Consequently we can also interpolate Image Primitives at any position along this feature :

$$\mathbf{e}_2, \mathbf{e}_3 \rightarrow I(s) \tag{2}$$

$s \in [0,1]$ being the position of the primitive along the feature, 0 being $\mathbf{e}_2$ and 1 being $\mathbf{e}_3$. We use that at our advantage to estimate the internal consistency of a Triplet : As $\mathbf{e}_1$ is linked to $\mathbf{e}_2$ and $\mathbf{e}_3$ then $T(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ means that the three primitives are manifestation of the same scene feature. Consequently, if we use $\mathbf{e}_2$ and $\mathbf{e}_2$ to interpolate a primitive at $\mathbf{e}_1$ position, the interpolated Image Primitive should be identiqual to the original one : consequently we reformulate our definition of a valid Triplet as follows. $T(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is a triplet if

1. links $l(\mathbf{e}_1, \mathbf{e}_2)$ and $l(\mathbf{e}_1, \mathbf{e}_3)$ exist.
2. $\mathbf{e}_2$ and $\mathbf{e}_3$ allow an interpolation I close enough to $\mathbf{e}_1$

5.2 Stereo Hypothesis Interpolation

Our use of the neighbourhood of a Image Primitive has allowed us to remove a number of stereo hypothesis. Yet we also need a mechanism to *add*, when the neighbourhood strongly hints towards a disparity. We simply extend the BSCE criterion to our new structure. A Triplet in the left image representation is a manifestation of a feature of the scene, and this feature should have a manifestation in the right image representation. Consequently, if the Image Primitives have a correspondence in the second image, then the Triplet should also be conserved through those correspondences. Consequently if we consider a triplet $T(\mathbf{e}_1^L, \mathbf{e}_2^L, \mathbf{e}_3^L)$ in the left image. If we also have the stereo hypothesis $s(\mathbf{e}_2^L) = \mathbf{e}_2^R$ and $s(\mathbf{e}_3^L) = \mathbf{e}_3^R$, then we can create a new stereo hypothesis

$$s_T(\mathbf{e}_3^L) = I(\mathbf{e}_2^R, \mathbf{e}_3^R) \tag{3}$$

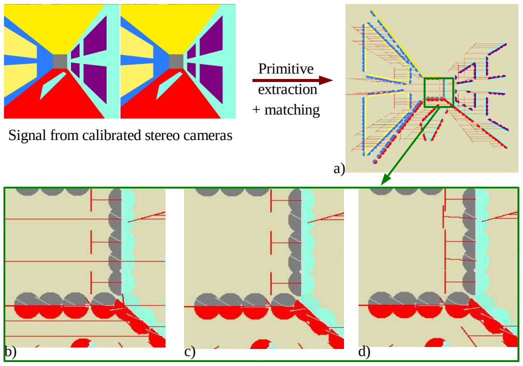
Signal from calibrated stereo cameras
Primitive extraction + matching
a)
b)
c)
d)

Fig. 4. Example on an artificial video sequence. a) shows the Image Primitives extracted from the left image, the red lines shows the position of their most likely stereo-hypothesis. The figures b), c), d) are zoomed on an interesting area, to be more readable. b) shows the original correspondences, using the multimodal criterion. c) shows the same area with a thresholding on the external confidence. Some obviously wrong correspondences disappear. Finally, d) shows after using the Triplet stereo-hypothesis interpolation. New hypothesis have been interpolated for primitives devoid of potential correspondence in the previous images. Those interpolated primitives allow a fuller reconstruction of those features of the scene — especially the gray–cyan boundary of the far surface.

6 Results and Conclusion

This combined approach of stereo–grouping of features has been applied to an artificial sequence — figure 4 — and to a real life scene recorded from a car — figure 5. Through our definition of the external confidence a large number of inconsistent stereo–hypothesis are being removed. Some of those would have been impossible to remove using purely local information, and some of them would have appeared as the best correspondence to local algorithms. Let us emphasize that those correspondences are from fundamentally correct primitives, and due to the natural repetitivity of scenes, and cannot be identified using purely local feature matching. On the other hand, the stereo–hypothesis interpolation through context allows to get a more complete set of correspondences, and ultimately a smoother and fuller reconstruction of 3D features. Effectively, the overall quality of our stereo–matching and of the following reconstruction is largely improved via the application of those two processes, only through local interaction. Also, the groups we defined through this process are inherently stereo ones. We think an iterative process joining the stereo and the grouping, to build larger entities can allow to remove irrelevant Image Primitives, to build groups more accurately using a richer constraint that the traditional collinearity, and correct stereo–assumptions. The resulting stereo–coupled features are then directly related to the relevant 3D features.

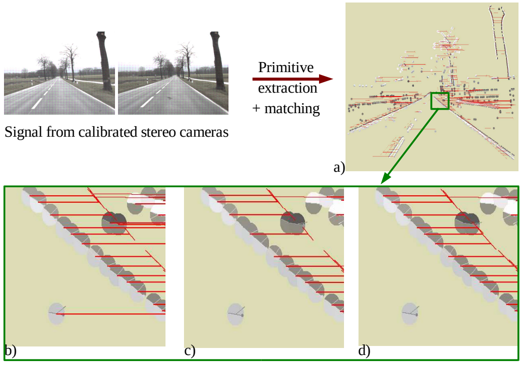

Fig. 5. Example of our processes on a real life video sequence of a driving scene. a) shows the Image Primitives extracted from the left image, the red lines shows the position of their most likely stereo-hypothesis. The figures b), c), d) are zoomed on an interesting area, to be more readable. b) shows the original correspondences, using the multimodal criterion. c) shows the same area with a thresholding on the external confidence. Some obviously wrong correspondences disappear. Finally, d) shows after using the Triplet stereo-hypothesis interpolation. New hypothesis have been interpolated for primitives devoid of potential correspondence in the previous images. Those interpolated primitives allow a fuller reconstruction of those features of the scene.

REFERENCES

[AS89]J. Aloimonos and D. Shulman. *Integration of Visual Modules — An extension of the Marr Paradigm*. Academic Press, London, 1989.

[FHH92]D. Field, A. Hayes, and R. Hess. Contour integration by the human visual system: Evidence for a local ”association field”. *Vision Research*, 33(2):173–193, 1992.

[KF04]N. Krüger and M. Felsberg. An explicit and compact coding of geometric and structural information applied to stereo matching. *accepted for Pattern Recognition Letters*, 2004.

[KJP02]N. Krüger, T. Jäger, and Ch. Perwass. Extraction of object representations from stereo imagesequences utilizing statistical and deterministic regularities in visual data. *DAGM Workshop on Cognitive Vision*, 2002.

[KLW03]N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *accepted for the Interdisciplinary Journal of Artificial Intelligence & the Simulation of Behavious, AISB Journal*, 1(4), 2003.

[KW02]N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *Network: Computation in Neural Systems*, 13:553–576, 2002.

[KW04]N. Krüger and F. Wörgötter. Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. *Advances in Imaging and Electron Physics*, 131, 2004.

[MLP$^+$96]Joseph L. Mundy, A. Liu, Nic Pillow, Andrew Zisserman, S. Abdallah, Sven Utcke, S. Nayar, and Charlie Rothwell. An experimental comparison of appearance and geometric model based recognition. In *Object Representation in Computer Vision*, pages 247–269, 1996.

[PK03]N. Pugeault and N. Krüger. Multi–modal matching applied to stereo. In *Proceedings of the BMVC 2003*, 2003.