AUTOMATIC DEPTH PROFILING OF 2D CINEMA- AND PHOTOGRAPHIC IMAGES

Dževdet Burazerović[1] (*dzevdet.burazerovic@philips.com*)
Patrick Vandewalle[2] (*patrick.vandewalle@philips.com*)
Robert-Paul Berretty[3] (*robert-paul.berretty@philips.com*)

Philips Research, The Netherlands

ABSTRACT

It is generally understood that interpreting depth from monoscopic images inherently induces the problem of object segmentation and context recognition. A common way to deal with this complexity is to focus on one aspect (e.g. segmentation) and largely constrain or skip the other (e.g. scene classification). In our approach, we practically reconcile the two paths. We start by defining several depth profiles, providing rough estimates for both background and foreground, while reflecting common photographic and cinematic practices and rules. The idea is to assign one or more of these profiles to an image – based on the output of multiple classifiers, and apply an image-adaptive bilateral filter to align the depth with object edges. From extensive trials, we conclude that robust classification and visual performance can be achieved with this approach on a variety of content.

Index Terms — Depth estimation, pattern recognition, image classification, image line pattern analysis.

1. INTRODUCTION

This paper presents an algorithm for automated estimation of depth from 2D photographic images, which aims to reconcile drawbacks of typical approaches. In essence, we seek a compromise between being generic about the source content and being precise when it comes to estimation of depth. The goal is to yield 'displayable' depth maps from diverse images, such that annoying artifacts are avoided and the depth effect possibly restrained.

1.1. Background and related work

It is generally understood that accurate interpretation of depth from arbitrary monocular images requires no less than human vision. Even when faithful segmentation of all structures in the scene can be automatically obtained, it sometimes takes high-level cognitive inference to establish the ordering of the different parts. The two aspects are also inherently intertwined: knowing the *scene context* can facilitate *object segmentation* and *recognition*, and vice-versa. Still, exact interaction of these two priors is not easily generalized; so, to avoid the chicken-and-egg problem, authors usually focus on one aspect and largely constrain or skip the other.

Some recent studies can be well distinguished along these lines. For instance, [1] assumes a fixed limited geometry: *ground*, *sky* and *vertical planes*, and uses this as prior to a learning algorithm that generates full segmentation and labeling of depth – which is done by connecting the so-called "super-pixels". With [2] it's the other way around: a classifier is trained to detect a larger number of "stages" – roughly describing depth layouts for the background, and the object segmentation is left aside. It is here assumed that the size and position of stage objects will be largely constrained by the stage. Some methods also operate in between, by engaging both image segmentation and classification, but less intricately [3]. Additional constraints are here usually imposed by detection of various *depth cues* (e.g. vanishing lines, defocus, etc.).

At the other end, some authors choose to stay implicit by inferring depth output directly from low-level image features; still, the trade-off between precision and generality remains as intrinsic as above. For instance, while [4] generate full depth maps with fidelity that largely varies with the type of source content, [5] compute only an average depth for the entire scene.

Roughly speaking, most algorithms operate in line with one of the philosophies outlined above. Those which are explicit about image segmentation are usually more geared towards the depth effect, but also artifacts. Conversely, shifting the focus to depth-cue or image classification by definition can offer only partial solutions.

1.2. Overview of the algorithm

Motivated by the above discussion, we opt for a balance between image/depth segmentation and classification. A way to concertize this is to classify both background *and* foreground, but without being too explicit or elaborate about either. The premise is that, by adopting smooth depth modeling and fewer target classes, we may reduce misclassifications and artifacts and still yield an appreciable depth effect. The essence of our approach is explained as follows:

- Background (foreground) is modeled via depth gradients, assuming several relative strengths and positions along the vertical (horizontal) direction.
- A depth profile may be composed from one or more of these gradients, based on the output of multiple classifiers operating on the source image.
- The final depth map is obtained by refining such a depth profile by an image-adaptive *bilateral filter* [6].

Figure 1 illustrates the different parts. The pre-processing connotes optional image down-scaling or edge-aligned smoothing [7], which will not affect the global depth profile, but may simplify its classification. This classification will be realized using multiple classifiers trained to detect depth gradients. Finally, the resulting depth profile will be processed by a bilateral filter that will look at local color differences in the source image for adaptation of its weights. Efficient cascading of such filter at multiple scales [8] can even yield a single depth shaping/up-scaling operation.

In the sequel, we shall first give some more background and motivation for our depth profiles, after which we shall elaborate the classification part. The last two sections will be devoted to our experiments and conclusions drawn from that.

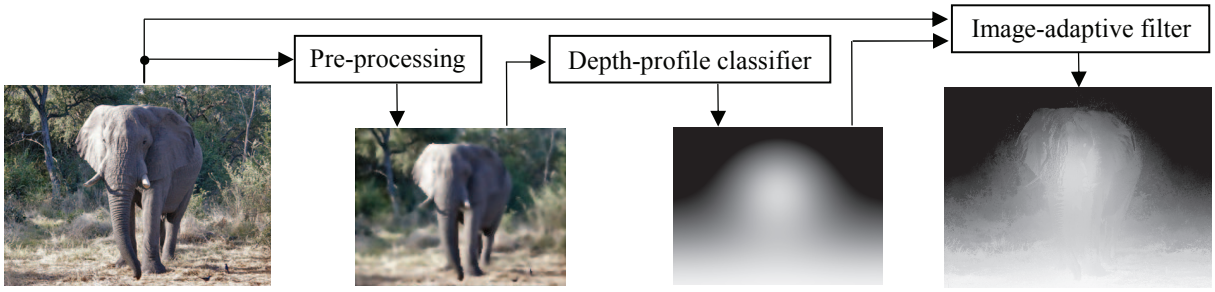

Figure 1: An overview of the algorithm. Notice the composite nature of the (manually tuned) profile and the depth-shaping action of the filter (may be less pronounced on a paper printout).

2. DEPTH-PROFILE CLASSIFICATION

Following the line from above, our goal is to trade some precision in defining the anatomy of depth profiles for robustness of their classification. Relying on the image-adaptive filter for fine-shaping of the depth map, we opt for a smaller number of global profiles.

2.1. Class definition

In vision research, the 'weak fusion model' [9] has been pointed as a close approximation of how human brain combines cues – i.e. by giving preference to those that deviate least from what was learned to be 'true'. An implication can be that the brain will often have an easier task in snubbing erroneous depth ordering than sharp depth misalignments. Therefore, we wish to concentrate on localization of significant visual transitions in the image, leaving their ordering to plausible heuristics or dedicated post-processing.

As a start, we consider classes such as: *centre, vertical* and *side*. A depth heuristics may then dictate that the former two respectively connote *foreground* and *top-bottom*, as is often entailed by the scenery and how people photograph objects of interest. The distinction of *left* and *right* within *side* is less tangible and may be treated as a separate task. Figure 2 gives few examples to illustrate all this. Notice the distinctive 'over-the-shoulder' close-up, which is a vastly used technique for filming dialogues in film!

Next, we consider hybrids, i.e. conjunctions of these classes, like e.g. *centre* $\wedge$ *vertical* in Figure 1. Such composites can be defined a priori, but also shaped afterwards, minding the confidences and correlations obtained from separate detections of their constituents. This will be explained in more detail in Section 2.4.

2.2. Feature extraction

It is intuitively clear that notable visual transitions and gradients in a true depth map can be well reflected in its *intensity profiles* – 1D signals obtained by accumulating the pixel-values along some line segments, e.g. columns and rows. The shape (trend, curvature and inflection points) of these profiles would likely suffice to qualify all classes of our interest. The problem, of course, is that we do not have the true depth maps but the source images instead.

Still, it can be verified that simple statistics, e.g. *mean* and *std*, of pixel rows and columns can yield distinctive 1D-profiles also from photographs. Such statistics can be extracted from multiple image planes and color spaces; however, mere *luma* will prove to yield ample discriminatory power for the classification problem at hand.

To quantify such a 1D-profile, we employ the well-known *DCT transform*. This will allow representing the profile as a sum of cosines with different frequencies, where the weight of each cosine is induced by the value of a corresponding *DCT coefficient*. This now leads to a simple concept: the larger the contribution of low-order DCT coefficients, the visually simpler is the image, and thus more likely it is to manifest notable depth gradients. Thus, we form our *feature set* by taking few initial (non-DC and normalized) DCT coefficients obtained by transforming the lower-order pixel-line statistic mentioned above. Figure 3 partly illustrates this.

(a) image (b) hand-drawn (c) depth profile

Figure 2: Examples of images incurring our basic classes: *vertical*, *centre* and *side*, including possible depth profiles to associate with.

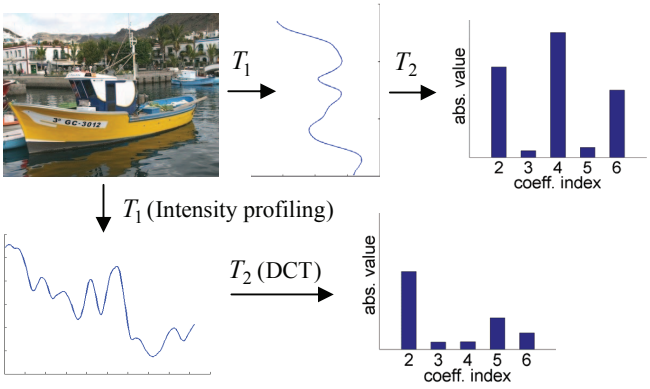

Figure 3. Feature extraction. Here, the profiling means *averaging* of pixel rows/columns of the *luma*, followed by *low-pass filtering*.

Additional Analysis. The main goal of the adopted DCT-based analysis was to enable a distinction of visual gradients with respect to their global course. For more gradient modulation, one should measure the 1D-profiles more precisely, e.g. by also locating their points of high curvature [10]. Still, one must realize that not all such points will designate true object/depth boundaries (see also Figure 3), which dictates caution concerning their a priori use.

2.3. Classification and decision fusion

From what was explained in Section 2.1, our approach conjuncts multiclass classification and information fusion. A suitable way to address both would be to engage multiple '1-1' (*one-against-one*) and '1-r' (*one-against-rest*) classifiers in *voting* [11]; dispersion of the votes could then suggest using only the winning class, or a hybrid of classes. However, in our case it is plausible that the posterior probabilities given by these classifiers will not deviate drastically from prior probabilities, due to fair ambiguity of our classes. Autonomous detection of each class and explicit handling of class confidences and correlations then seem convenient.

Component Classifiers. In line with the above, we refrain from wide optimizing of separate classifiers, except for letting each use a different subset of our features. We back this by the notion that the number (≤ 18) and orthogonality of our features will fit the optimal working range of several popular classifiers methods, if we provide enough training samples [12]. We adopt *quadratic discriminant analysis* as a suitable classifier, and *backward elimination* as a way of selecting the best features for each class. Detection of i-th class, ω_i, will then be incurred via (1), where $\mathbf{x}$ denotes a n_i-dimensional feature vector and $\{\mathbf{A}_i, \mathbf{B}_i, C_i\}$ the model parameters derived (via *maximum likelihood estimation*) from a training set for that class.

$$S_i \equiv S(\omega_i \mid \mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{B}_i\mathbf{x} + C_i; \quad d_i = \begin{cases} 1, & S_i \geq 0 \\ 0, & S_i < 0 \end{cases} \tag{1}$$

To read decisions as probabilities, we use the sigmoid mapping (2), where α_i can be tuned to give P_i the assumed Gaussian-like density.

$$P_i \equiv P(\omega_i \mid \mathbf{x}) = \left(1 + \exp(-\alpha_i S_i)\right)^{-1}; \quad 0 < P_i < 1 \tag{2}$$

Decision Fusion. A suitable way to account for class correlations is to treat classifiers' decisions (1) as a *Behavior-Knowledge Space* [13]. This will allow mapping of each out of 2^N possible N-tuples of 0's and 1's – as given by N classifiers, to a specific class or set of classes. This mapping is to be learned by arranging the N-tuples and class-counts obtained from many images, via a look-up table.
Yet, for fractional class assignment, we shall prefer using the class probabilities (2) and some known models for correlated random events. We are mostly interested in class *conjunctions*, and one formula to estimate them is given by Frank's model [14] as:

$$P_{ij} \equiv P(\omega_i \wedge \omega_j) = \log_s\left(1 + (s^{P_i} - 1)(s^{P_j} - 1)/(s - 1)\right) \tag{3}$$

where $s = \tan(\pi(1 - \rho_{ij})/4)$ and ρ_{ij} is the class correlation, satisfying: $\rho_{ij} \notin \{-1, 0, 1\}$. Admittedly, this is just one of many models we could use; though, its being a copula has some general advantages [14].
Since P_i, P_j in (3) will come from classifiers, it is fair to estimate ρ_{ij} from classifiers' decisions, e.g. on a large random set of M images:

$$\rho_{ij} = \frac{1}{M}\sum_{t=1}^{M} q_{ij}(t); \quad q_{ij} = \begin{cases} 1, & d_i = d_j \\ 0, & d_i \neq d_j \end{cases} \tag{4}$$

With the probabilities of separate classes (2) and their conjunctions (3) in place, it is now possible to construct analytical and rule-based procedures for weighting the different classes as patches of a depth map. One such procedure is explained below.
Note also that we shall omit (3) with salient conjunctions, which can be better treatable as self-reliant classes, using (2). In fact, we could use both procedures in parallel and average their results.

2.4. Shaping the profiles

We now describe our way of translating the class probabilities (2), (3) into a depth profile. We start by defining a general formula for shaping such a profile, in (5), where D denotes depth and (x, y) the normalized pixel coordinates. The 2D Gaussian is mainly intended to allow placing a 'blob' of certain size anywhere in the image, and D_1 for creating smooth horizontal or vertical gradients. The bottom expression is used for scaling and additional modulation.
Figure 4 demonstrates this. Note that the left and centre profile differ in that for $\{x_0, y_0, y_1, b_2\}$ they use $\{0.25, 0.5, 0.8, 0.6\}$ and $\{0.15, 0.6, 0.4, 1.5\}$, respectively. Comparing the shapes, we can see the role of these parameters in relation to our classes. Since x_0 specifies the horizontal shift of the 'blob', it may be controlled by the conjunction of *centre* and *side*. In turn, $\{y_0, y_1, b_2\}$ regulate the height and tail of the profile, which should depend on conjunctions of the two horizontal classes with *vertical*. These mappings and other parameters in (5) allow various class-oriented pre-settings, which we here omit. The profiles from Fig. 1, 2 and 4 show some typical designs. Important is that, in case of simultaneous detection of multiple classes, the associated profiles will be summed with weights proportional to the respective class probabilities.

$$\begin{aligned} D_0(x,y) &= e^{-\left(\frac{(x-x_0)^2}{a_1} + \frac{(y-y_0)^2}{a_2}\right)} \\ D_1(x,y) &= b_1(x - x_1) + b_2(y - y_1) \\ D &= gain \cdot \max(0, w_0 D_0 + w_1 D_1) \end{aligned} \tag{5}$$

side ∧ centre *side ∧ vertical* *two sides*

Figure 4. Examples of depth profiles created using (5) and their association with our target classes and their conjunctions.

3. EXPERIMENTS AND RESULTS

Classification Performance. The classifiers' training and evaluation were done with some 500 photos and movie frames, with the latter group contributing by two thirds. The set reflected significant content diversity, including portrayals of humans and animals, landscapes and decors, with homogeneous and complex backgrounds. Quite expectedly, *medium-views* and *close-ups* were prominent in the movies, as opposed to *long* and *full-views* [15] typical of outdoor photography. By adoption of *stratified holdout* and 40% test set, 250-400 images were effectively used per class.
Table 1 shows the average detection rates for the most deserving classes; see also Figures 2, 4. Not surprisingly, the rates decline as we go from basic to more composite classes. The class correlations

are shown in Table 2 and also conform to general expectations. For instance, reckon that *vertical* will conjunct with *centre* or *side* if the foreground, designated by the latter two, occludes some far background. Also, *s2* and *centre* will be exactly the opposites in case of homogeneous background, etc. It is thus clear that some degree of confusion between our classes will be unavoidable when aiming for large content diversity, as we have. This in a way also justifies the probabilistic class assignment explained above.

Table 1. Detection rates of the component classifiers

Class	Precision	Recall [%]
vertical, steep (v1)	85.45	79.11
vertical, gradual (v2)	67.39	68.31
side (s)	73.47	70.85
centre (c)	71.90	68.50
side-and-*centre (sc)*	63.01	58.44
two sides (s2)	64.89	57.46

Table 2. Class correlations (from classifiers' decisions)

	v1	*v2*	*s*	*c*	*sc*	*s2*
v1						
v2	0.34					
s	0.18	0.21				
c	0.12	0.16	0.24			
sc	0.03	0.04	0.10	0.11		
s2	0.15	0.14	0.06	0.19	0.06	

Visual Performance. This part was done with yet a different set of some 80 images, for which accurate depth maps had been drawn by hand. See also Figure 2. The set included all the assorted types of source content and camera shots mentioned above (though not in equal amounts). The depth profiles were built according to classes from Table 1 and the procedure explained in Sec. 2.4. By lack of adequate heuristics, *side* was designated as *left* or *right* randomly. The *image+depth* rendering was done on a 42" lenticular display. The visual quality was rated with reference to *manual annotations* and in comparison to [16]. We summoned our ratings via a *mean opinion score*, as shown in Figure 5. The curve corresponding to our method shows a logical trend, while its favorable placement can be explained by the diversity of the image content. So, while [16] did produce preferable results in scenes that complied with its own assumptions, it more often created annoying artifacts – due to incoherent depth layering, i.e. object segmentation. In contrast, our misclassifications were mostly manifested as a more dispersed loss of sharpness. (Imagine a slight misalignment of the depth profile in Figure 1 and the ability of the image-adaptive filter to 'diffuse' it.)

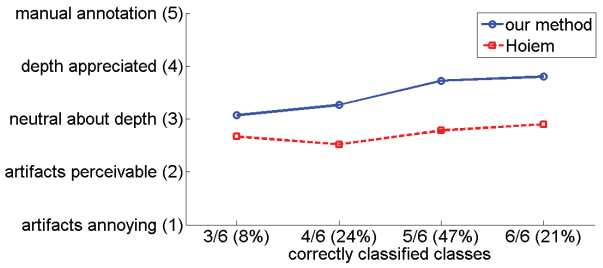

Figure 5. MOS of 3 viewers, obtained with a diverse test set. Each sample connotes the *average* of all images where a certain number of classes from Table 1 were correctly labeled by our method. The percentages tell how often this occurred in total. Note that 'Hoiem' refers to software implementing [16], which is available online.

4. CONCLUSIONS

In this paper, we propose a method for estimating depth from 2D photos and film-frames, which effectively reconciles the normally separated depth classification and segmentation. This is achieved by classifying a smaller number of smooth depth profiles and using an image-adaptive filter for their refinement. In practical terms, such approach is meant to offer a worthy balance between artifacts and the depth effect on diverse image content. Our tests and results confirm a fulfillment of that. Lastly, we believe our classification is also usable beyond the application described in the paper, as it can be extended to other tasks of image- and video-content analysis.

REFERENCES

[1] D. Hoiem, A.A. Efros, M. Herbert, "Automatic Photo Pop-up", *ACM SIGGRAPH*, 2005.

[2] V. Nedovic, A.W.M. Smeulders, A. Redert, J.-M. Geusebroek, "Depth Information by Stage Classification," *Proc. IEEE ICCV,* pp. 1-8, 14-21 Oct. 2007.

[3] S. Battiato, S. Curti, M. La Cascia, M. Tortora, E. Scordato, "Depth-map Generation by Image Classification", *Proc. SPIE,* Vol. 5302, 95, 2004.

[4] A. Saxena, S. H. Chung, A. Y. Ng, "Learning depth from single monocular images", *NIPS*, 2005.

[5] A. Torralba, A. Oliva, "Depth Estimation from Image Structure", *IEEE Trans. on PAMI*, Vol. 4, No. 9, 2002.

[6] C. Tomasi, R. Manduchi, "Bilateral Filtering for Gray and Color Images", *Proc. ICCV*, Bombay, India, 1998.

[7] P. Perona, J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. PAMI*, V. 12, no. 7, pp. 629-639, 1990.

[8] A.K. Riemens, O.P. Gangwal, B. Barenbrug, R-P.M. Berretty, "Multi-step joint bilateral depth upsampling", in SPIE Vol. 7257: *Proc. VCIP*, 2009.

[9] M. Landy, L. Maloney, E. Johnston, M. Young, "Measurement and modeling of depth cue combination: In defense of weak fusion", *Vision Research*, Vol. 35, pp. 389–412, 1995.

[10] H-H. Liu, J-H. Twu, S-J. Wang, "Image Representation Using Curvature Information in Intensity Profiles", *Proc. IEEE ICIP*, Vol. 2, pp. 720 – 723, Oct. 1997.

[11] D.M.J. Tax, R.P.W. Duin, "Using two-class classifiers for multiclass classification", *Proc. Int. Conf. on Pattern Recognition,* Vol. 2, pp. 124-127, 2002.

[12] J. Hua, et al. "Optimal number of features as a function of sample size for various classification rules", *Bioinformatics*, Vol. 21, No. 8, pp. 1509-1515, 2005.

[13] Y.S. Huang, C.Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals", *IEEE Trans. on PAMI*, 17:90-93, 1995.

[14] S. Ferson et. al, "Dependence in Dempster-Shafer theory and probability bounds analysis", *Technical Report SAND 2004-3072*, Sandia National Laboratory, 2004.

[15] D. Arijon, "Grammar of the Film Language", Communication ArtsBooks, Hastings House, NY, 1976.

[16] D. Hoiem, A. Stein, A. Efros, M. Hebert, "Recovering Occlusion Boundaries from a Single Image", *ICCV*, 2007.