

OBJCUT for Face Detection

Jonathan Rihan, Pushmeet Kohli, and Philip H.S. Torr
{jon.rihan, pushmeet.kohli, philiptorr}@brookes.ac.uk

Department of Computing
Oxford Brookes University, UK
<http://cms.brookes.ac.uk/research/visiongroup/>

Abstract. This paper proposes a novel, simple and efficient method for face segmentation which works by coupling face detection and segmentation in a single framework. We use the OBJCUT [1] formulation that allows for a smooth combination of object detection and Markov Random Field for segmentation, to produce a real-time face segmentation. It should be noted that our algorithm is extremely efficient and runs in real time.

1 Introduction

Object detection and segmentation are important problems of computer vision and have numerous commercial applications such as pedestrian detection, surveillance and gesture recognition. Image segmentation has been an extremely active area of research in recent years [2–5, 1, 6]. In particular segmentation of the face is of great interest due to such applications as Windows Messenger[©] [7, 8].

Until recently the only reliable method for performing segmentation in real time was blue screening. This method imposes strict restrictions on the input data and can only be used for certain specific applications. Recently Kolmogorov *et al.* [9] proposed a robust method for extracting foreground and background layers of a scene from a stereo image pair. Their system ran in real time and used two carefully calibrated cameras for performing segmentation. These cameras were used to obtain disparity information about the scene which was later used in segmenting the scene into foreground and background layers. Although they obtained excellent segmentation results, the need for two calibrated cameras was a drawback of their system.

Shape priors for Segmentation: An orthogonal approach for solving the segmentation problem robustly has been the use of prior knowledge about the object to be segmented. In recent years a number of papers have successfully tried to couple MRFs used for modelling the image segmentation problem with information about the nature and shape of the object to be segmented [3, 4, 1, 10]. The primary challenge in these systems is that of ascertaining what would be a good choice for a prior on the shape. This is because the shape (and pose) of objects in the real world vary with time. To obtain a good shape prior then, there is a need to localize the object in the image and also infer its pose, both of which are extremely difficult problems in themselves.

Kumar *et al.* [1] proposed a solution to these problems by matching a set of exemplars for different parts of the object on to the image. Using these matches they generate a shape model for the object. They model the segmentation problem by combining

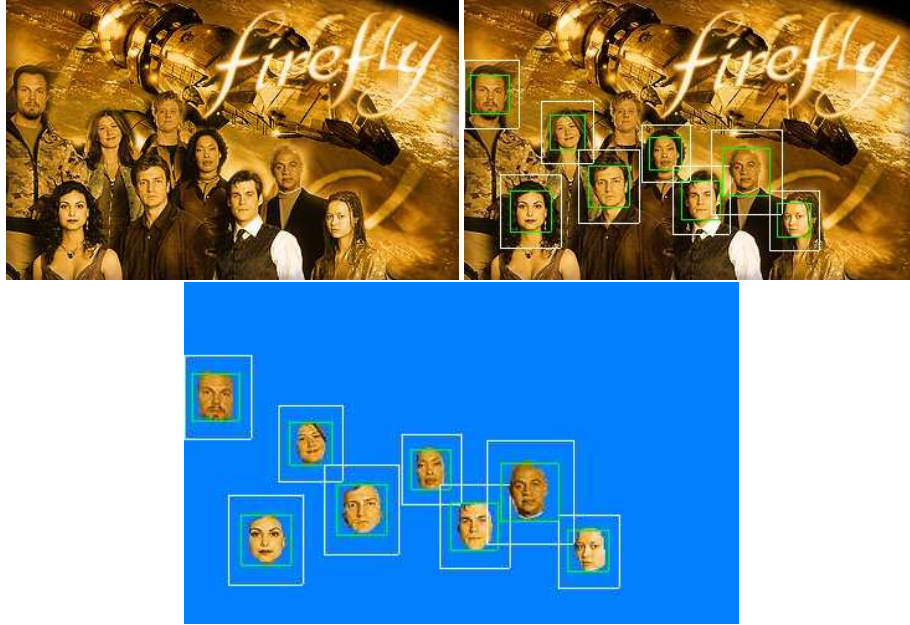


Fig. 1. Real Time Face Segmentation using a face detections. The first image on the first row shows the original image. The second image shows the face detection results. The image on the second row shows the segmentation obtained by using shape priors generated using the detection and localization results.

MRFs with layered pictorial structures (LPS) which provide them with a realistic shape prior described by a set of latent shape parameters. A lot of effort has to be spent to learn the exemplars for different parts of the LPS model.

In their work on simultaneous segmentation and 3D pose estimation of humans, Bray *et al.* [3] proposed the use of a simple 3D stick-man model as a shape prior. Instead of matching exemplars for individual parts of the object, their method followed an iterative algorithm for pose inference and segmentation whose aim was to find the pose corresponding to the human segmentation having the maximum probability (or least energy). Their iterative algorithm was made efficient using the dynamic graph cut algorithm [5]. Their work had the important message that *rough shape priors were sufficient to obtain accurate segmentation results*. This is an important observation which will be exploited in our work to obtain an accurate segmentation of the face.

Coupling Face Detection and Segmentation: In the methods described above the computational problem is that of localizing the object in the image and inferring its pose. Once a rough estimate of the object pose is obtained, the segmentation can be computed extremely efficiently using graph cuts [2, 5, 11–13]. In this paper we show how an off the shelf face-detector such as the one described in [14] can be coupled with graph cut

based segmentation to give accurate segmentation and improved face detection results in real time.

The key idea of this paper is that face localization estimates in an image (obtained from any generic face detector) can be used to generate a rough shape energy. These energies can then be incorporated in to a discriminative MRF framework to obtain robust and accurate face segmentation results as shown in Figure 1. This method is an example of the OBJCUT paradigm for an unarticulated object. We define an uncertainty measure corresponding to each face detection which is based on the energy associated with the face segmentation. It is shown how this uncertainty measure might be used to filter out false face detections thus improving the face detection accuracy.

Organization of the Paper: This paper proposes a method for face segmentation which works by coupling the problems of face detection and segmentation in a single framework. Our method is extremely efficient and runs in real time¹. The key novelties of the paper include:

- A framework for coupling face detection and segmentation problems together.
- A method for generating rough shape energies from face detection results.
- An uncertainty measure for face segmentation results which can be used to identify and prune false detections.

A summary of the paper follows. In the next section, we briefly discuss the methods for robust face detection and image segmentation. In section 3, we describe how a rough shape energy can be generated using localization results obtained from any face detection algorithm. The procedure for integration of this shape energy in the segmentation framework is given in the same section along with details of the uncertainty measure associated with each face segmentation. We conclude by listing some ideas for future work in section 4.

2 Preliminaries

In this section we give a brief description of the methods used for face detection and image segmentation.

2.1 Face Detection and Localization

Given an image, the aim of a face detection system is to detect the presence of all human faces in the image and to give rough estimates of the positions of all such detected faces. In this paper we use the face detection method proposed by Viola and Jones [14]. This method is extremely efficient and has been shown to give good detection accuracy. A brief description of the algorithm is given next.

The Viola Jones face detector works on features which are similar to Haar filters. The computation of these features is done at multiple scales and is made efficient by using an image representation called the *integral image* [14]. After these features have

¹ We have developed a system which uses a single camera and runs in real time.

been extracted, the algorithm constructs a set of classifiers using AdaBoost [15]. Once constructed, successively more complex classifiers are combined in a cascade structure. This dramatically increases the speed of the detector by focussing attention on promising regions of the image. The output of the face detector is a set of rectangular windows in the image where a face has been detected. We will assume that each detection window W_i is parameterized by a vector $\theta_i = \{c_i^x, c_i^y, w_i, h_i\}$ where (c_i^x, c_i^y) is the centre of the detection window and w_i and h_i are its width and height respectively.

2.2 Image Segmentation

Given a vector $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ where each y_i represents the colour of the pixel i of an image having n pixels, the image segmentation problem is to find the value of the vector $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ where each x_i represents the label which the pixel i is assigned. Each x_i takes values from the label set $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$. Here the label set \mathcal{L} consists of only two labels i.e. ‘face’ and ‘not face’. The posterior probability for \mathbf{x} given \mathbf{y} can be written as:

$$\Pr(\mathbf{x}|\mathbf{y}) = \frac{\Pr(\mathbf{y}|\mathbf{x}) \Pr(\mathbf{x})}{\Pr(\mathbf{y})} \propto \Pr(\mathbf{y}|\mathbf{x}) \Pr(\mathbf{x}). \quad (1)$$

We define the energy $E(\mathbf{x})$ of a labelling \mathbf{x} as:

$$E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{y}) + \text{constant} = \phi(\mathbf{x}, \mathbf{y}) + \psi(\mathbf{x}) + \text{constant}, \quad (2)$$

where $\phi(\mathbf{x}, \mathbf{y}) = -\log \Pr(\mathbf{y}|\mathbf{x})$ and $\psi(\mathbf{x}) = -\log \Pr(\mathbf{x})$. Given an energy function $E(\mathbf{x})$, the most probable or maximum a posterior (MAP) segmentation solution \mathbf{x}^* can be found as:

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} E(\mathbf{x}). \quad (3)$$

It is typical to formulate the segmentation problem in terms of a Discriminative Markov Random Field [16]. In this framework the likelihood $\phi(\mathbf{x}, \mathbf{y})$ and prior terms $\psi(\mathbf{x})$ of the energy function can be decomposed into unary and pairwise potential functions. In particular this is the contrast dependent MRF [2, 5] with energy:

$$E(\mathbf{x}) = \sum_i (\phi(x_i, \mathbf{y}) + \psi(x_i)) + \sum_{(i,j) \in N} (\phi(x_i, x_j, \mathbf{y}) + \psi(x_i, x_j)) + \text{constant}, \quad (4)$$

where N is the neighbourhood system defining the MRF. Typically a 4 or 8 neighbourhood system is used for image segmentation which implies each pixel is connected with 4 or 8 pixels in the graphical model respectively.

Colour and Contrast based Segmentation: The unary likelihood terms $\phi(x_i, \mathbf{y})$ of the energy function are computed using the colour distributions for the different segments in the image [2, 1]. For our experiments we built the colour appearance models for the face/background using the pixels lying inside/outside the detection window obtained from the face detector. The pairwise likelihood term $\phi(x_i, x_j, \mathbf{y})$ of the energy function is called the *contrast term* and is *discontinuity preserving* in the sense that it encourages

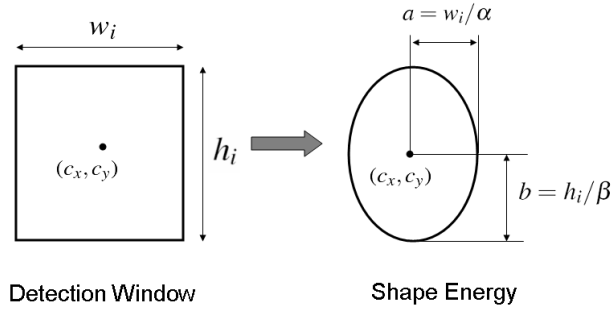


Fig. 2. Generating the face shape energy. The figure shows how a localization result from the face detection stage is used to generate a rough shape energy for the face.

pixels having dissimilar colours to take different labels (see [2, 1] for more details). This term takes the form:

$$\phi(x_i, x_j, \mathbf{y}) = \begin{cases} \gamma(i, j) & \text{if } x_i \neq x_j \\ 0 & \text{if } x_i = x_j. \end{cases} \quad (5)$$

where $\gamma(i, j) = \exp\left(\frac{-g^2(i, j)}{2\sigma^2}\right) \frac{1}{\text{dist}(i, j)}$. Here $g^2(i, j)$ measures the difference in the RGB values of pixels i and j and $\text{dist}(i, j)$ gives the spatial distance between i and j .

The pairwise prior terms $\psi(x_i, x_j)$ are defined in terms of a generalized Potts model as:

$$\psi(x_i, x_j) = \begin{cases} K_{ij} & \text{if } x_i \neq x_j, \\ 0 & \text{if } x_i = x_j. \end{cases} \quad (6)$$

and encourage neighbouring pixels in the image² to take the same label thus resulting in smoothness in the segmentation solution. In most methods, the value of the unary prior term $\psi(x_i)$ is fixed to a constant. This is equivalent to assuming a uniform prior and does not effect the solution. In the next section we will show how a shape prior derived from a face detection result can be incorporated in the image segmentation framework.

3 Integrating Face Detection and Segmentation

Having given a brief overview of image segmentation and face detection methods, we now show how we couple these two methods in a single framework. Following the OBJCUT paradigm, we start by describing the face energy and then show how it is incorporated in the MRF framework.

The face shape energy: In their work on segmentation and 3D pose estimation of humans, Bray *et al.* [3] show that rough and simple shape energies are adequate to obtain accurate segmentation results. Following their example we use a simple elliptical

² Pixels i and j are neighbours if $(i, j) \in N$

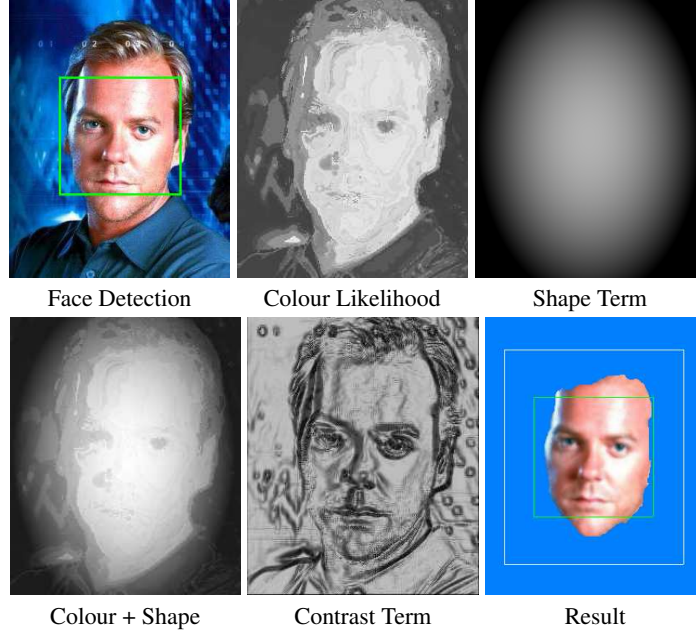


Fig. 3. Different terms of the shape-prior + MRF energy function. The figure shows the different terms of the energy function for a particular face detection and the corresponding image segmentation obtained.

model for the shape energy for a human face. The model is parameterized in terms of four parameters: the ellipse centre coordinates (c_x, c_y) , the semi-minor axis a and the semi-major b (assuming $a < b$). The values of these parameters are computed from the parameters $\theta_k = \{c_k^x, c_k^y, w_k, h_k\}$ of the detection window k obtained from face detector as: $c_x = c_k^x$, $c_y = c_k^y$, $a = w_k/\alpha$ and $b = h_k/\beta$. The values of α and β used in our experiments were set to 2.5 and 2.0 respectively, however these can be computed iteratively in a manner similar to [3]. A detection window and the corresponding shape prior are shown in figure 2.

3.1 Incorporating the Shape Energy

For each face detection k , we create a shape energy Θ_k as described above. This energy is integrated in the MRF framework described in section 2.2 using the unary terms $\psi(x_i)$ as:

$$\psi(x_i) = \lambda(x_i|\Theta_k) = -\log p(x_i|\Theta_k) \quad (7)$$

where we define $p(x_i|\Theta_k)$ as:

$$p(x_i = \text{'face'}|\Theta_k) = \frac{1}{1 + \exp(\mu * (\frac{(cx_i - c_x^k)^2}{a^{k^2}} + \frac{(cy_i - c_y^k)^2}{(b^k)^2} - 1))} \quad (8)$$

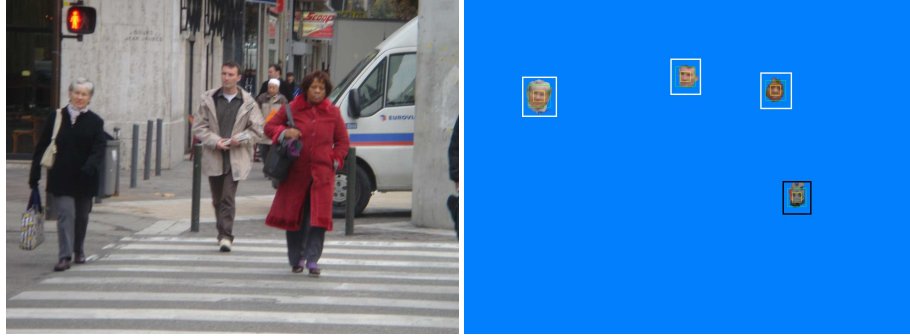


Fig. 4. The figure shows an image from the INRIA pedestrian data set. After running our algorithm, we obtain four face segmentations, one of which (the one bounded by a black square) is a false detection. The energy-per-pixel values obtained for the true detections were 74, 82 and 83 while that for the false detection was 87. As you can see the energy of false detection is significantly higher than that of the true detections, and can be used to detect and remove it.

$$\text{and } p(x_i = \text{'background'} | \Theta_k) = 1 - p(x_i = \text{'face'} | \Theta_k) \quad (9)$$

where cx_i and cy_i are the x and y coordinates of the pixel i , $\{c_x^k, c_y^k, a^k, b^k\}$ are parameters of the shape energy Θ_k , and the parameter μ determines how the strength of the shape energy term varies with the distance from the ellipse boundary. The different terms of the energy function and the corresponding segmentation for a particular image are shown in figure 3.

Once the energy function $E(\mathbf{x})$ has been formulated, the most probable segmentation solution \mathbf{x}^* defined in equation (3) can be found by computing the solution of the max-flow problem over the *energy equivalent* graph [13]. The complexity of the max-flow algorithm increases with the number of variables involved in the energy function. Recall that the number of random variables is equal to the number of pixels in the image to be segmented. Even for a moderate sized image the number of pixels is in the range of 10^5 to 10^6 . This makes the max-flow computation quite time consuming. To overcome this problem we only consider pixels which lie in a window W_k whose dimensions are double of those of the original detection window obtained from the face detector. As pixels outside this window are unlikely to belong to the face (due to the shape term $\psi(x_i)$) we set them to the background. The energy function for each face detection k now becomes:

$$E_k(\mathbf{x}) = \sum_{i \in W_k} \phi(x_i, \mathbf{y}) + \psi(x_i | \Theta_k) + \sum_{j \in W_k, (i,j) \in N} \phi(x_i, x_j, \mathbf{y}) + \psi(x_i, x_j) + \text{constant}, \quad (10)$$

This energy is then minimized using graph cuts to find the face segmentation \mathbf{x}_k^* for each detection k .

Pruning false detections: The energy $E(\mathbf{x}')$ of any segmentation solution \mathbf{x}' is the negative log of the probability, and can be viewed as a measure of how uncertain that

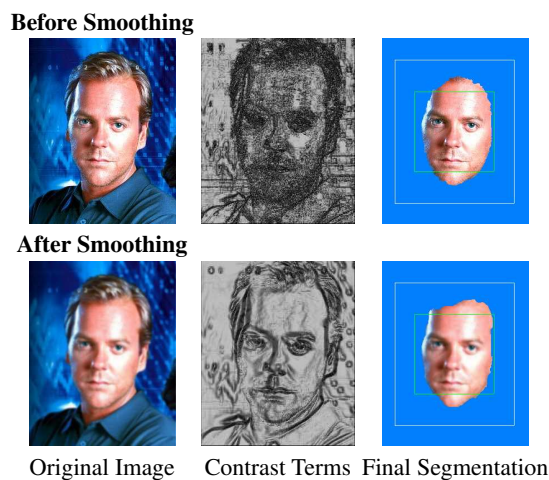


Fig. 5. Effect of smoothing on the contrast term and the final segmentation. The images on the first row correspond to the original noisy image. The images on the second row are obtained after smoothing the image.

solution is. The higher the energy of a segmentation, the lower the probability that it is a good segmentation. Intuitively, if the face detection given by the detector is correct, then the resulting segmentation obtained from our method should have high probability and hence have low energy compared to the case of a false detections (as can be seen in figure 4). This characteristic of the energy of the segmentation solution can be used to prune false face detections. Alternatively, if the number of people P in the scene is known, then we can choose the top P detections according to the segmentation energy.

3.2 Implementation and Experimental Results

We tested our algorithm on a number of images containing faces. Some detection and segmentation results are shown in figure 6. The time taken for segmenting out the faces is of the order of tens of milliseconds. We also implemented a real time system for frontal face detection and segmentation. The system is capable of running at roughly 15 frames per second on images of 320x240 resolution.

Handling Noisy Images: The contrast term of the energy function might become quite bad in noisy images. To avoid this we smooth the image before the computation of this term. The result of this procedure are shown in figure 5.

4 Conclusion and Future Work

In this paper we presented a method for face segmentation which combines face detection and segmentation into a single framework. Our method runs in real time and gives accurate segmentation and improved face detection results.

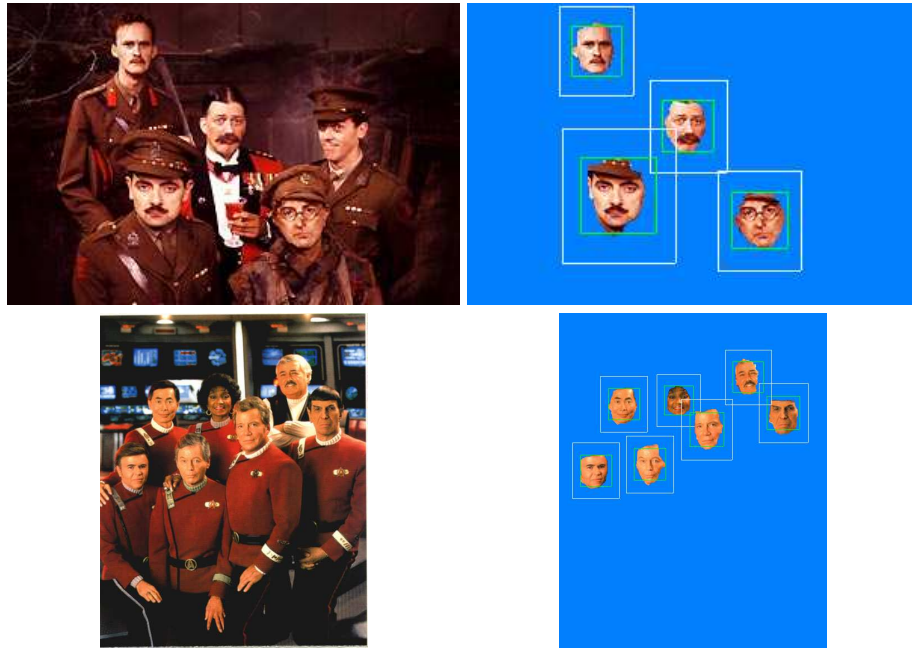


Fig. 6. Some face detection and segmentation results obtained from our algorithm.

While segmenting image frames of a video, the use of knowledge of the correct face detections in the previous frames in eliminating errors in the current image frame needs to be explored. Another area for future research is the idea of efficient selective refinement of the shape energy. This procedure could successively refine the shape energy to obtain good segmentations in complicated scenarios. It should be noted that such a procedure could be performed using dynamic graph cuts [5] which would make it computationally efficient.

References

1. Kumar, M., Torr, P., Zisserman, A.: OBJ CUT. In: CVPR. Volume I. (2005) 18–25
2. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: ICCV. Volume I. (2001) 105–112
3. Bray, M., Kohli, P., Torr, P.: PoseCut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph cuts. In: ECCV. (2006) 642–655
4. Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: CVPR. Volume I. (2005) 755–762
5. Kohli, P., Torr, P.: Efficiently solving dynamic markov random fields using graph cuts. In: ICCV. (2005)
6. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. (2004) 309–314
7. Criminisi, A., G, C., Blake, A., V, K.: Bilayer segmentation of live video. In: IEEE Computer Vision and Pattern Recognition (CVPR). (2006)

8. Sun, I., Zhang, W., Tang, X., Shum, H.: Background cut. In: ECCV. (2006)
9. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Bi-layer segmentation of binocular stereo video. In: CVPR (2). (2005) 407–414
10. Huang, R., Pavlovic, V., Metaxas, D.: A graphical model framework for coupling mrfs and deformable models. In: CVPR. Volume II. (2004) 739–746
11. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI **26** (2004) 1124–1137
12. Greig, D., Porteous, B., A., S.: Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society **2** (1989) 271–279
13. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? In: ECCV. Volume III. (2002) 65 ff.
14. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision (2004)
15. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational Learning Theory: Eurocolt 95. (1995) 23–37
16. Kumar, S., Hebert, M.: Discriminative fields for modeling spatial dependencies in natural images. In: NIPS. (2003)