

# i3DPost Deliverable 4.3: Semantic Labelling of Point Clouds (Public Summary)

## I. INTRODUCTION

One of the aims of i3DPost is to provide a structured and semantically rich description of multi-view videos and 3D scenes. This will be achieved through research that will be performed in two tasks. More specifically task WP4T3 ("Semantic labelling of 3D point clouds") will provide a first labelling the scene elements, by identifying humans and important objects (e.g. props) and tagging them with person identity and object class information, respectively. Subsequently, task WP4T5 ("Semantic description of human activity for retrieval") will enrich the descriptions that will be derived in WP4T3 with semantic annotation with respect to the state (e.g. facial expressions) and actions of humans.

In more detail, at the first stage humans in the scene will be identified. This will be achieved through human body detection and pose estimation techniques. Moreover, identification and semantic labelling of the scene's objects will be attempted. In a way analogous to assigning scene objects to one of the predefined object categories, appropriate person recognition techniques will be used to recognize and label the scene actors with their identities.

The result of task "Semantic labelling of 3D point clouds" will be a semantic description of scene elements (such as identity and body pose of persons as well as class/category of objects) in each time instance of a multi-view or 3D scene. This semantic information will subsequently be used (along with the description of actor actions and state derived in task "Semantic description of human activity for retrieval") in a proper metadata format for the intelligent manipulation of the content and its efficient and semantic organization, search and retrieval. A rough schematic representation of the aims of task WP4T3 is provided in Figure 1.

This deliverable describes the research activities that took place and the preliminary results that were obtained within Task WP4T3 in the first 12 months of the project. In this first stage of the project, the focus was mainly on techniques operating on either single-view or multiple-view data. The deliverable consists of 4 sections, each dealing with a different subtopic, namely human detection and body pose estimation, human recognition using facial information, object detection and categorization and finally a method for 3D object simplification for faster registration that can be used along with template-based body detection and pose estimation or activity recognition techniques.

## II. MULTI-VIEW HUMAN DETECTION AND BODY POSE ESTIMATION

A novel method for human body detection and pose estimation has been devised. The method will be used to detect humans in sequences captured by multiple synchronized and calibrated cameras. Moreover

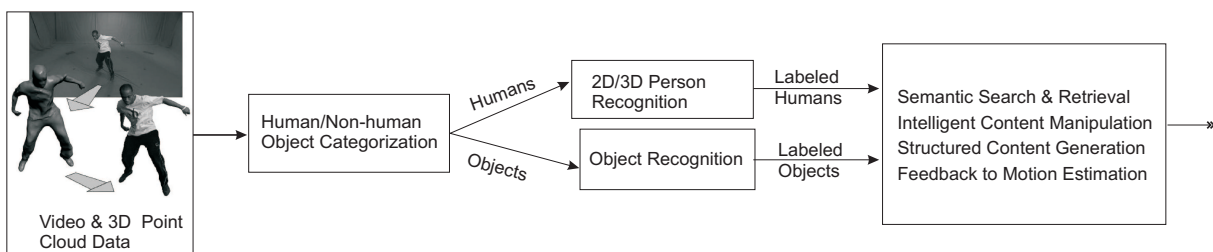


Fig. 1. A schematic representation of the aims of WP4T3.

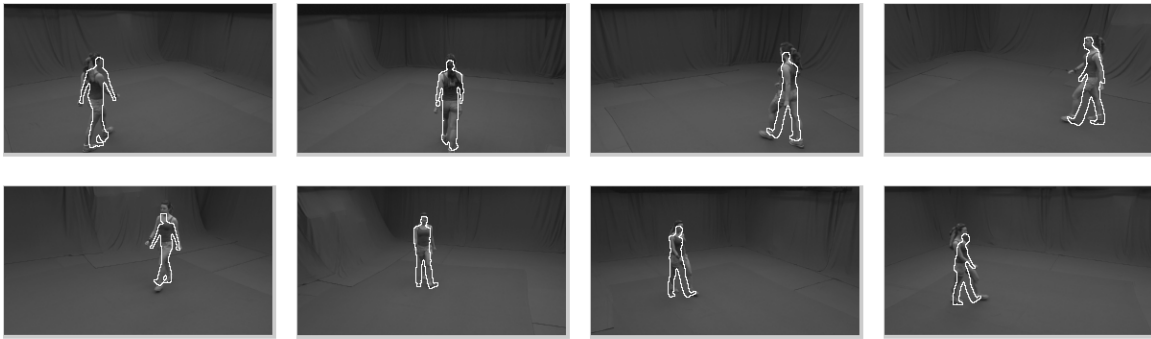


Fig. 2. Silhouettes matched in the 8 different views of the same time instance of a walking sequence.

the body poses discovered in consecutive multi-view frames might be subsequently used along with temporal information for activity recognition or for improving matting and/or motion estimation.

The proposed method uses a novel self-balanced binary search tree data structure for the storage and fast shape matching of human silhouette templates. The tree stores silhouette templates of a human performing actions (currently only walking), generated by rendering from a number of view angles actions of synthetic 3D human models. Two types of nodes exist: leaf (template) nodes that store the templates and internal nodes that don't contain any actual template information and are used to determine the search path to the leaf nodes where the actual data is stored. Matching is performed using the modified Hausdorff distance.

Using this shape tree we wish to detect the presence of humans inside each multiview frame, locate them and estimate their body pose in each view as well as the 3D pose. The procedure used to achieve the above is as follows: We first use the binary tree to perform human detection on each individual frame (view). Knowing the parameters of each camera, we can then get an estimate of the location of the detected human in 3D space. We then move the search to the 3D space and project 3D locations that are candidates for containing a human back to the individual frames. For each particular location in the 3D space, all individual frames cast votes on a) which 3D pose they observe and b) from which view angle they observe it. After making a final decision on the pose and the orientation of the person in the 3D space, a final matching and refinement is performed.

The proposed method is currently evaluated with respect to its ability to estimate the body pose using multi-view data filmed within the project at the University of Surrey. The data were captured using a convergent eight high definition cameras setup. Promising preliminary results have been obtained so far. Figure 2 illustrates a sample result obtained at the end of the pose estimation process. The proposed multi-view method was also compared with a single view variant of this method that utilizes the same tree and a similar methodology. The multi-view method produced better and more consistent preliminary results than the single view one.

### III. PERSON RECOGNITION

Within the topic of person recognition, three lines of activity have been pursued during the first 12 months of the project.

The first activity dealt with a kernel-based method for face verification from images that is based on a novel optimization criterion giving rise to class specific discriminant projections. Although face verification is a 2-class problem (clients and impostors classes), many verification approaches treat the feature extraction part of this task within a  $N$ -class framework which results in the extraction of non person specific features. For example, when Fisher's Linear Discriminant Analysis (FLDA) is applied in a  $N$ -class problem formulation, the discovered transformation is not person specific. Unfortunately, when FLDA is applied in the more proper (for the problem at hand) 2-class formulation, it generates transformed data of very small dimension (1-D or 2-D), which is a very strict limitation. The aim of this

activity was to remedy these limitations, exploit human face individuality and take into consideration the non-linearity of the problem.

This led to a novel kernel discriminant algorithm that was named Class-Specific Kernel Discriminant Analysis (CSKDA). The main features of this algorithm are that a) a 2-class formulation for the face verification is adopted, b) the algorithm involves a new optimization criterion for person specific discriminant feature extraction, c) kernel techniques are applied to capture the nonlinearity of facial image distribution under different facial poses and expressions and d) with appropriate training the method can operate in a view-independent way.

The most important characteristic of the proposed algorithm is that it aims to find a discriminant projection in the Hilbert space (that results from the application of the kernels). This projection is such that the sum of distances of the projected impostor feature vectors from the mean feature vector  $\bar{\rho}$  of the client class is maximized whereas the sum of distances of the projected client feature vectors from  $\bar{\rho}$  is minimized. The algorithm derives person-specific projections and a different training should be performed for each of the  $N$  "i-th client-vs-impostors" 2-class problems. Unlike FLDA, that generates in such a case 1-D or 2-D projections, the dimensions of the subspace generated by CSKDA are proportional to the number of training images.

The proposed method has been used for single view frontal face verification on the ORL and Yale facial databases and for view-independent verification on the XM2VTS and UMIST databases. For view-independent face verification in the XM2VTS and UMIST databases, views of the same person captured from different view angles have been used for training whereas a single facial image taken from an arbitrary view was used for testing. The experiments showed that the proposed method outperforms KPCA and multiclass KFDA. For example, in the view-independent experiments in the XM2VTS database the proposed method achieved an Equal Error Rate (EER) of 3.3% whereas KFDA achieved 6.8% and KPCA 10.2%

It should be noted that although the proposed method deals with the problem of face verification, it can easily be adapted to solve the problem of face recognition which is of interest to i3DPost. This work led to the following publication:

*S. Zafeirou, G. Goudelis, A. Tefas, N. Nikolaidis, I. Pitas, Motivating Class-Specific nonlinear projections for single view and multiview face verification, IEEE International Conference on Image Processing (ICIP 2008), October 12-15, San Diego, CA, U.S.A.*

The second line of activity focused on a method for the extraction of discriminant features for use in face-related classification algorithms. The method, called Projected Gradient Discriminant Non-negative Matrix Factorization (PGDNMF) is an extension of the Discriminant Non-negative Matrix Factorization (DNMF) algorithm, which in turn has been derived from the Non-negative Matrix Factorization (NMF) algorithm. The method has been tested in the problem of face verification but features derived from facial images can be also used for face recognition and facial expression recognition.

The NMF algorithm which is the basis of both DNMF and PGDNMF aims at decomposing a set of images (in our case, facial images) into a linear combination of a set of basis images:

$$\mathbf{X} \approx \mathbf{Z}\mathbf{H}. \quad (1)$$

where  $\mathbf{X}$  is the vector containing all images to be decomposed,  $\mathbf{Z}$  is the matrix containing the basis images and  $\mathbf{H}$  is the decomposition coefficients vector. Both  $\mathbf{H}$  and  $\mathbf{Z}$  contain non-negative elements and are evaluated through an iterative procedure that tries to minimize the decomposition error.

The DNMF is a supervised algorithm that utilizes information regarding the classes of the training images in order to impose discriminant constraints on the decomposition. In more detail, apart from minimizing the decomposition error, it tries to minimize over  $\mathbf{Z}$ ,  $\mathbf{H}$  the scatter of features within a class and maximize the scatter between classes. This is done by minimizing the following cost function subject to non-negativity constraints for  $\mathbf{H}$  and  $\mathbf{Z}$ :

$$D_d(\mathbf{X}||\mathbf{ZH}) = D(\mathbf{X}||\mathbf{ZH}) + \gamma\text{tr}[\mathbf{S}_w] - \delta\text{tr}[\mathbf{S}_b] \quad (2)$$

where  $D(\mathbf{X}||\mathbf{ZH}) = \sum_j KL(\mathbf{x}_j||\mathbf{Z}\mathbf{h}_j)$  is the sum of Kullback-Leibler (KL) divergences for all images in the set (decomposition error). The matrices  $\mathbf{S}_w$ ,  $\mathbf{S}_b$  are the within-class and the between-class scatter matrices respectively. Unfortunately, the update rules of DNMF only ensure the non-increasing behavior during the minimization of (2). On the contrary, the proposed PGDNMF algorithm guarantees the convergence of the algorithm to a stationary point. Moreover in DNMF, the discriminant constraints are applied on the decomposition coefficients  $\mathbf{H}$ , and not on the actual features  $\tilde{x}$  used in the subsequent classification i.e., those derived from the projection of the facial images  $x$  on the basis images. However, in the proposed method, the discriminant constraints are applied directly on the actual features  $\tilde{x}$ . For achieving the above, a new, modified optimization problem is defined and solved using projected gradients in order to guarantee that the limit point will be stationary.

The proposed algorithm has been initially applied on the problem of face verification along with 2-class SVM classifiers. Experiments were conducted on the XM2VTS face database using a standard evaluation protocol. The method was compared to a number of other feature extraction/decomposition approaches namely NMF, LNMF, DNMF, Class Specific DNMF, PCA and PCA plus LDA. The best Equal Error Rate achieved by PGDNMF was equal to 2% which was better than all other methods apart from that of PCA+LDA (FisherFaces) that achieved a slightly better EER or 1.7%

This work led to the following publication:

*I. Kotsia, S. Zafeiriou and I. Pitas, Discriminant Non-negative Matrix Factorization and Projected Gradients for Frontal Face Verification. 1st COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008), 7-9 May 2008, Roskilde, Denmark*

The third line of activity within the person recognition topic of Task WP4T3 dealt with 3D head pose estimation in single-view video sequences acquired by a static uncalibrated camera. This work does not deal directly with person recognition, however, the head pose estimates derived from the proposed method can be used in different ways in order to assist the face recognition task as well as the facial expression recognition task. One such way is to select the frames that depict the person in a neutral/frontal (or nearly frontal) pose and use them to recognize the face or the facial expression through related algorithms that require a frontal view. Another way would be to use head orientation data along with a cluster of classifiers trained to recognize faces or facial expressions from different view angles.

In the algorithm that has been devised, head pose estimation is performed using in an incremental way a structure from motion (SfM)/self-calibration technique. The usual mode of operation of such algorithms is the following: given a set of photographs of a static rigid object captured by a moving camera the algorithm estimates the 3D structure of the object and derives the camera parameters for each photograph. Since there is an obvious duality between a moving camera that captures a static object and a static camera that captures a moving object, in our case we use such an algorithm in a reciprocal way: given a number of video frames of a moving object (head) captured by a static camera, estimate the 3D structure of the object and the camera calibration parameters (location, orientation and intrinsic parameters) for each frame and use the latter to straightforwardly derive the orientation parameters of the moving head.

In more detail, face detection and tracking are used in order to obtain the facial bounding boxes in a video. Then, using successive triads of frames that overlap by two frames, Scale Invariant Image Transform (SIFT) image keypoints are detected and matched (with the assistance of Random Sample Consensus (RANSAC) for the rejection of outlying matches), and camera calibration is performed using the algorithm proposed in [1]. Assuming that the head in the first frame of the sequence is in a neutral/frontal orientation the derived camera orientation can be used to easily estimate the orientation of the head in each frame. Since each triad of frames involves two already processed frames along with a new one, the estimates

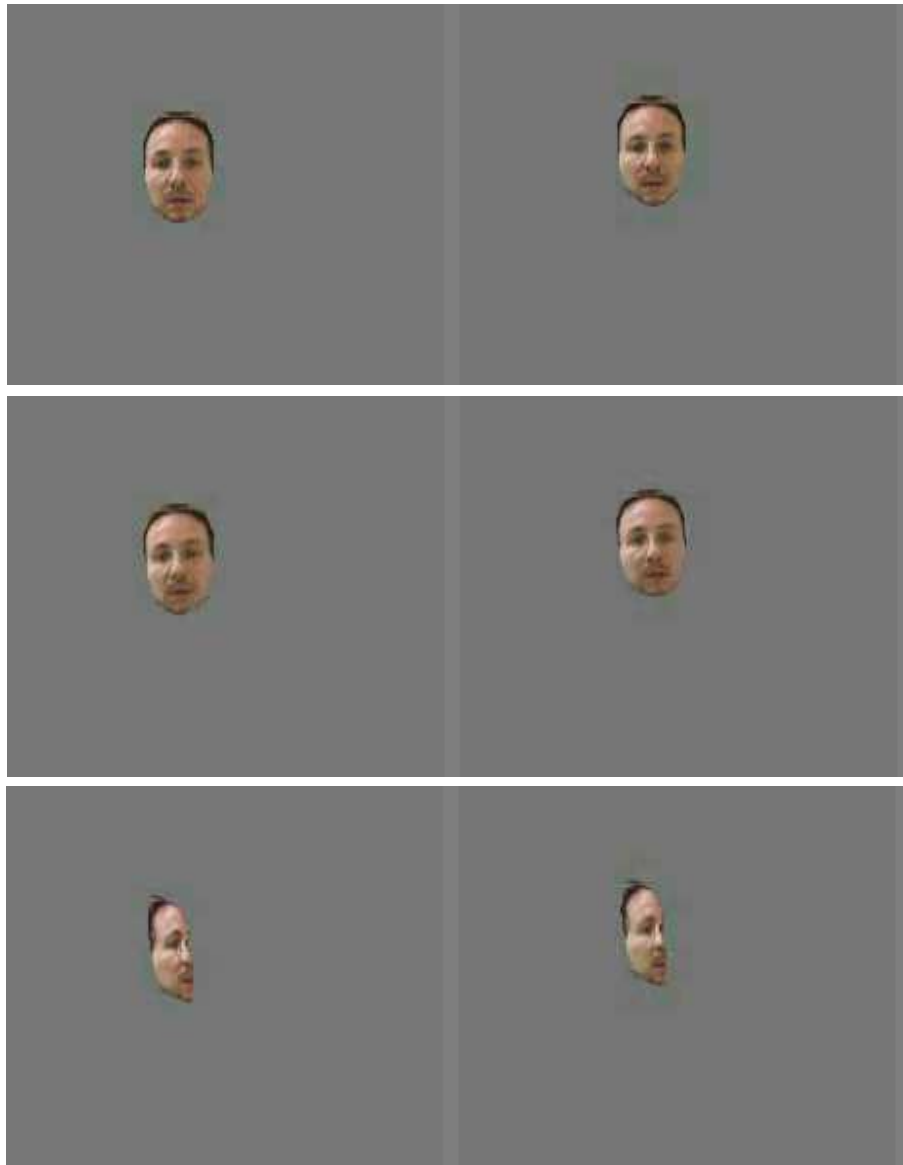


Fig. 3. Examples of results obtained from the proposed head pose estimation algorithm.

from the application of the algorithm in the previous triad of frames are used in the so-called bundle adjustment step of the SfM algorithm in order to improve the results.

The proposed method was used to estimate the 3D head pose on the IDIAP video database [2] that contains videos depicting meeting and office environments as well as head pose ground truth in the form of pan, tilt and roll angles for each frame of the video sequences. The metrics used for the evaluation of the proposed head pose estimation algorithm were the error (absolute difference) in degrees between the ground truth and the estimated value for pan, tilt and roll angles. Moreover, the head pose defines a vector in the 3D space which indicates where the head is pointing at. The angle (DE) between the pointing vector defined by the head pose ground truth and the pose estimated by the proposed system was also used as a pose estimation error measure of the face. The proposed algorithm was found to be able to estimate the 3D head pose with satisfactory accuracy. The mean error in the pointing vector (DE) for the proposed algorithm is  $18.9^\circ$  while the error obtained by the best algorithm in [2] is  $21,3^\circ$ . The average absolute error between estimated and ground truth angles was found to be  $14.50^\circ$  for roll,  $18.40^\circ$  for pan and  $11.20^\circ$  for tilt. Some results of the head pose estimation algorithm are depicted in Figure 3.

Each image of this figure depicts on the left part the textured Canide facial model oriented according to the ground truth head pose angles and on the right part the same model oriented according to the angles estimated by the proposed algorithm.

#### IV. OBJECT DETECTION AND RECOGNITION

An approach for object detection, and object (or object class) recognition has been investigated. The approach is based on local features derived with the SIFT algorithm, and a database of pre-rendered photorealistic images of the types of objects we are interested in. Given an image that may contain objects of interest, matches are found in the database using local features and then detection, class recognition and identity recognition are performed by a voting procedure assisted by outlier rejection.

In more detail, we used a 3D modelling and rendering software package in order to photorealistically render a number of example 3D models in a variety of postures. At this preliminary investigative stage, we limited the scope of our database to furniture, and more specifically to kitchen chairs. For the selection and representation of our feature points we have selected the Scale Invariant Feature Transform (SIFT) [3] that evaluates for each point a 128-dimensional feature vector. For each image in the database we evaluate a set of keypoints, characterized by a SIFT descriptor and tagged with its position in the image as well as its associated scale and rotation. In order to perform object detection and recognition in an input image we first perform SIFT keypoint detection in this image. Then, for each keypoint  $p_i$  found, similar keypoints  $p_{jk}$  are matched in the database images (indexed by  $j$ ) based on the dissimilarity  $\Delta x_{ijk} = \|x_i - x_{jk}\|$  of their SIFT descriptors. The speed of this computationally demanding search and matching step is improved using Locality Sensitive Hashing (LSH) [4]. This procedure provides, for every keypoint  $p_i$  in the query image, a set of matching keypoints  $p_{jk}$  in the database images along with their corresponding distances  $\Delta x_{ijk}$ . We are thus able to estimate the overall  $\widetilde{\Delta x}_j$ , i.e. the dissimilarity of the query image from the  $j$ -th database image by using each retrieved/matched  $p_{jk}$  as a weighted vote for this image.

However, some of the matches established with the above procedure might be erroneous and should be removed. Detection of erroneous matches is done by finding, using an iterative algorithm, the pairs that do not follow the global affine transformation model that relates the query image with a database image. Thus, the result of the algorithm is a set of graphics images from the database that are ordered with respect to their similarity  $1/\widetilde{\Delta x}_j$  with the query image. From this set we can choose the image with the highest similarity  $\widetilde{\Delta x}_j$  (and thus decide that the query image depicts the object displayed in this database image) or perform a voting procedure where, for example, each database image votes for the object  $\theta_j$  it depicts.

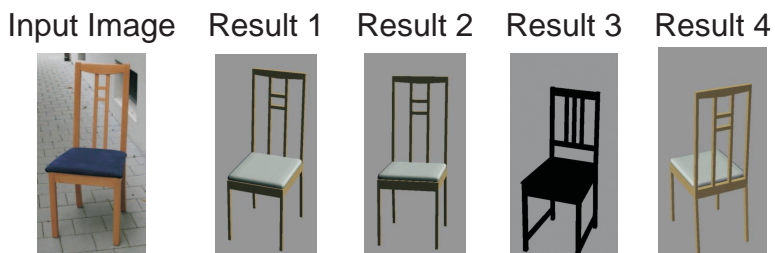


Fig. 4. Example of the results of our algorithm. The first image is the input image, and the others are the best matches in the database in order of similarity.

We have performed a small set of preliminary experiments to verify the performance of our algorithm. The results in general were satisfactory, as illustrated in Figure 4. In the next period we plan to make further improvements and extensions to the method and run a comprehensive set of experiments.

#### V. OBJECT SIMPLIFICATION FOR FASTER REGISTRATION

Work on object simplification for faster registration has been also performed. This work is based on background research that resulted in a method for simplifying sets of points in a way that is optimal

with respect to a subsequent registration or matching of such sets. In more detail, the method aims at simplifying sets of points so that the registration of the reduced sets results in minimal increase of the registration error compared to the registration of the full sets. The proposed simplification algorithm belongs to the Centroidal Voronoi Tessellation framework [5], [6]. Within i3DPost, the method has been adapted successfully for the simplification of volumetric data and 3D point clouds (such as vertices defining a 3D surface) that are to be registered through the Iterative Closest Point (ICP) algorithm and experiments were conducted. The method can be used in i3DPost for the speedup of techniques that require data registration and matching such as 3D-template based body detection and pose estimation or activity recognition techniques. The experiments conducted on human body 3D surface models and volumetric data showed that indeed the simplification method has minimal effect on the ICP registration accuracy and that it decreases substantially the registration computational complexity.

This work resulted in the following publication:

A.Hajdu, P.Verés, A. Tanács, I. Pitas, "Simplification of objects for faster human body posture estimation and clinical registration", *Voronoi-2008 Workshop, in conjunction with Numerical Geometry, Grid Generation and Scientific Computing (NUMGRID2008), June 10-13, 2008, Moscow, Russian Federation.*

## VI. CONCLUSIONS

The research activities that took place within Task "Semantic labelling of point clouds) in the first 12 months of the project have been described in this deliverable. The results obtained so far in all activities are very promising. The resulting semantic description of scene elements, such as identity and body pose of persons as well as class of objects, will subsequently be used (along with the description derived in task WP5T4) for the intelligent manipulation of the content.

## REFERENCES

- [1] M. Pollefeys and L. V. Gool, "A stratified approach to metric self-calibration," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997, p. 407418.
- [2] S. Ba and J.-M. Odobez, "Evaluation of multiple cues head pose estimation algorithms in natural environments," in *IEEE International Conference on Multimedia and Expo (ICME)*, Amsterdam, December 2005.
- [3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.
- [4] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 518–529.
- [5] Q. Du, M. Gunzburger, and L. Ju, "Constrained centroidal voronoi tessellations on general surfaces," *SIAM J. Scientific Comp.*, vol. 24, pp. 1488–1506, 2003.
- [6] A. Hajdu and I. Pitas, "Optimal approach for fast object-template matching," *IEEE Trans. on Image Processing*, vol. 16, no. 8, pp. 2048–2057, 2007.