Image Guided Phase Unwrapping for Real–Time 3D–Scanning

Thilo Borgmann, Michael Tok and Thomas Sikora
Communication Systems Group
Technische Universität Berlin

***Abstract*—3D–reconstructions produced by active 3D–scanning systems based on structured light can achieve high accuracy reconstructions of the scene surfaces. Structured light algorithms based on phase measuring triangulation (PMT) utilize phase–shifted sinusoidal patterns projected into the scene for a precise determination of correspondencies. The number of patterns used for that purpose may vary depending on the design of the algorithm.**

No matter how many patterns are required, all of these algorithms suffer from the acquisition time needed to record all patterns sequentially. In case of a dynamic scene the sequential acquisition of images lead to the capture of dynamic objects in different poses which in turn result in erroneous reconstructions depending on the object's velocity. Our goal is to achieve a more robust result during dynamic scene capture as well as better scene reconstruction rate. Two novel approaches are presented to reduce the amount of required patterns for a high–accuracy 3D–reconstruction. This is achieved by incorporating passive matching techniques in the phase–unwrapping stage of the algorithm, allowing to drop one half of the sinusoidal patterns.

I. Introduction

The three–dimensional presentation of television programs and movies have become more and more common during the last years. Current setups for movie theaters as well as home entertainment are capable of presenting the content in a stereoscopic manner by utilizing two or more views of the scenery. Presenting different views of a scenery for each eye of the spectator results in a three-dimensional impression of the scenery by the visual exploitation of parallax. Along with these immersive capabilities comes the need for high–quality scene acquisition to produce suitable input for such systems [1][2].

3D–scanning systems based on structured light are feasible candidates for such high–quality scene acquisition whenever the illumination of the scene with a given intensity pattern is applicable. Possible settings for these scanners range from small objects up to volumes of several cubic meters. Even with such high–volume reconstructions very detailed representations of the scenery can be acquired [3].

The reconstruction process of such systems requires the projection of several well–known intensity patterns. Thus, capturing real–world scenes featuring dynamically moving objects becomes a very challenging task for such scanners due to the time needed for sequential projection and recording of these patterns. A dynamic object changes it's pose during illumination and in turn distorts the corresponding reconstruction [4]. Our main goal is to make the reconstruction less prone to motion artifacts caused by dynamic objects. Our approaches are also offering the possibility to raise the rate of reconstruction. We effectively reduce the number of patterns required while preserving a detailed and accurate high–quality reconstruction.

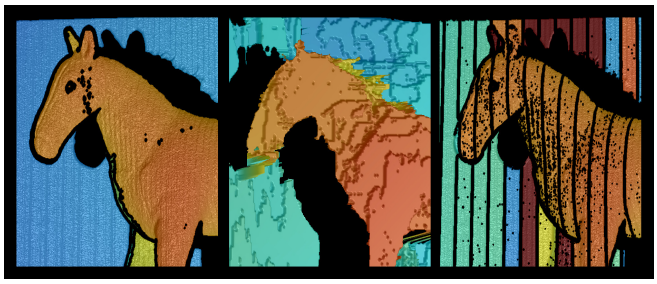

Fig. 1. 3D–reconstructions. From left to right: PMT, PCF, SPU.

In the following sections we will give a brief overview of related work, describe our approaches for reducing the number of required patterns, evaluate the achieved results and close with the conclusion and discussion of further work. A first impression of the achieved results is given in figure 1, showing the reconstructions of the basic algorithm [3] as well as both approaches presented in this paper.

II. Related Work

Image–based 3D–reconstruction is usually divided into active and passive reconstruction methods [5][6][7]. Nowadays there are several commercial systems available of both categories that allow for out–of–the–box acquisition of 3D–geometry. Binocular or even trifocal passive stereo cameras like the Point Grey Bumblebee cameras are good examples of ready–to–use products while the Microsoft Kinect or the ASUS Xtion are amongst available active stereo cameras. In general, the active variants can achieve much more detailed and accurate 3D–information than their passive counterparts which are also more prone to erroneous behavior because of homogeneous areas or repetitive patterns and alike.

Furthermore, the passive reconstruction methods are categorized into local and global methods. Global methods which are optimizing the estimated reconstruction using the full images provided by the sensor usually achieve most accurate passive reconstructions. Local methods that rely only on a defined surrounding area within the provided images generally

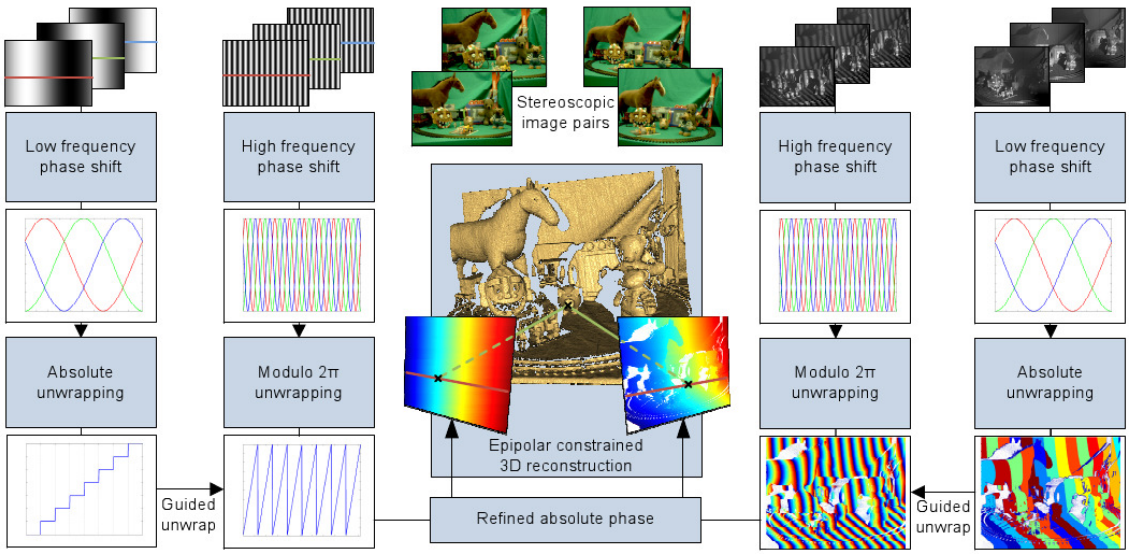

Fig. 2. Schematic diagram of the PMT algorithm [3].

compute less accurate results but are much faster, especially on modern GPGPU capable hardware. A comprehensive summary of passive algorithms of both types is given in [5]. For the scope of this work we make use of local passive reconstruction techniques for the sake of fast computation which is required for a real–time acquisition system we are dealing with.

From the family of active reconstruction techniques we relate to the class of structured light methods using phase measuring triangulation (PMT) [8][6][7]. The different phase–shifted patterns are projected into the scene and the 3D–reconstruction is computed by measuring the deformations within the pattern that is captured by the cameras. A good overview of 3D–reconstruction algorithms based on structured light is given in [9].

III. Image Guided Phase Unwrapping

In this work we base our approaches to reduce the amount of required patterns on the PMT method presented in [3]. We give a short overview of this method before describing our approaches to modify the phase unwrapping in detail.

As presented in [3], this method projects six phase–shifted sinusoidal patterns onto the scene for 3D–reconstruction. There are three high–frequency and three low–frequency patterns projected. For the high–frequency triplet the sinusoidal intensity ramp is wrapped thirty–two times from the left to the right of the image. For the low–frequency triplet exactly one wave is shown throughout the whole width of the image. The phase is shifted by 60 degrees for each pattern of both triplets, resulting in six distinct patterns.

Both triplets are projected sequentially onto the scene and captured by a synchronized monochromatic high–speed camera. Based on the resulting six images showing the distorted patterns captured from another point of view, as well as the six undistorted images used for projection, the absolute phase Φ within the camera and projector is computed as outlined in figure 2. The absolute phase Φ is computed for each pixel coordinate x using the number of wraps N and the modulo 2π unwrapped phases of the high–frequency and low–frequency pattern Φ'_h and Φ'_l, respectively:

$$\Phi(x) = \frac{\Phi'_h(x) + \lfloor N\Phi'_l(x) + 0.5 \rfloor}{N} \tag{1}$$

The modulo 2π phase $\Phi'_{h,l}(x)$ is computed depending on the three phase–shifted intensity patterns $p_{1,2,3}(x)$ of a given frequency triplet:

$$\Phi'_{h,l}(x) = \frac{arctan\left(\frac{\sqrt{3}(p_1(x)-p_3(x))}{2p_2(x)-p_1(x)-p_3(x)}\right)}{2\pi} \tag{2}$$

For the scope of this work, we utilized one 3D–scanning unit of the complete setup presented in [3]. This unit consists of a projector for illumination, a monochrome high–speed camera for phase acquisition as well as two color cameras for stereo matching used in both approached described in sections III-A and III-B. This setup is shown in figure 3.

In order to reconstruct three–dimensional world coordinates for a given pixel of the distorted image, the corresponding pixel in the undistorted image has to be found. Having

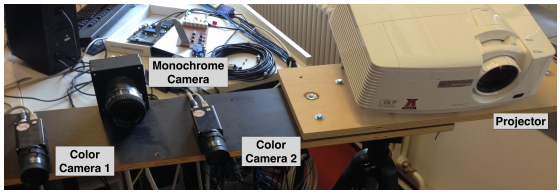

Fig. 3. Hardware configuration of the 3D–scanning unit.

Fig. 4. Color image of the scene and PMT based 3D–reconstruction.

computed the absolute phase Φ for both, the distorted images of the monochromatic camera as well the undistorted projected images, this correspondence can easily be found by searching for the same intensity or phase value along the epipolar line given by the calibrated geometry of the projector and camera setup [10]. This search yields the correspondence in pixel coordinates for both images that can be transformed into three–dimensional world coordinates using the calibration information. Then, two lines through the geometrical center of the respective camera or projector as well as the three–dimensional world coordinate of the corresponding pixel are determined. The intersection of both lines then results in the triangulated three–dimensional world coordinate observed through the corresponding pixels.

Computing these three–dimensional world coordinates for each pixel in the distorted image that has been illuminated by the projected patterns results in a dense high–quality 3D–reconstruction of the scene like shown in figure 4. For a more detailed description of this method, please refer to [3].

In the following we present two approaches to reduce the amount of required sinusoidal patterns of the PST method described. Both approached try to determine the absolute phase Φ using the three high–frequency patterns only, avoiding projection of the low–frequency pattern. First, we describe a bottom–up approach in section III-A trying to estimate the correspondance directly by incorporating the modulo 2π phase Φ'_h into the cost function of a passive reconstruction method. Second, we describe a top–down approach in section III-B trying to replace the calculation of $\lfloor N\Phi'_l(x)+0.5 \rfloor$ in equation (2) by estimating the corresponding segment of $\Phi'_l(x)$ based on a precomputed depth estimation. If the absolute phase can be determined successfully by these approaches, the image acquisition time needed for a single reconstruction is reduced by a factor of two and the low–frequency patterns can be skipped. A shorter acquisition time reduces the negative impact on the reconstruction result induced by the movement of a dynamic object. Also, less patterns to be projected allow for a higher rate of reconstruction.

A. Plane–Sweeping Based On Phase Cost Function (PCF)

Next to the high–speed monochromatic cameras and the projectors, the active 3D–reconstruction system also captures the scene using four color cameras that are also geometrically calibrated within the system. The acquired color images in [3] are used for texturing the reconstructed scene by using projective texturing techniques [11]. In order to drop the low–frequency patterns we incorporate a passive 3D–reconstruction method presented in [12], a method that is designed for a plane–sweep based reconstruction using multiple color images. For the first approach, this method is adapted so that the unwrapped high–frequency phase images are also taken into account for determining the pixel correspondencies.

Like the underlying active 3D–reconstruction, the passive plane–sweep method in [12] also requires a calibrated setup of cameras in order to determine three–dimensional world coordinates observed by corresponding pixels. Using this geometric information, a virtual plane is constructed representing a hypothetical planar reconstruction of the scene. Applying projective texturing [11], the images of the corresponding color cameras are projected onto that plane and blended together. Whenever the virtual plane intersects with the actual geometry of the scene, the projected color images blend into a locally undistorted version of the actual visible geometry [13]. The virtual plane featuring the blended textures is then projected into another camera which is a suitable reference as it features a second image that is composed by the real projection of the scene. The visual difference can be expressed by a given local aggregated cost metric between the blended hypothesis and the undistorted color image of the reference camera. The virtual plane is swept through a predefined volume yielding many different blended versions of the scene geometry composed by the color images. For each pixel in the image plane of the reference camera the minimum visual difference during the plane–sweep determines the pixel correspondencies between the composing color cameras required for triangulation.

The local visual cost of a plane hypothesis is computed by the locally aggregated pixel–wise cost function using the blended color image and the reference image. For the pixel–wise cost $C(x)$ computation, we calculate the AD–CENSUS measure for the pixel coordinate x using of the absolute pixel differences (C_{AD}) and the Hamming distance within a local window (C_{census}), described in detail in [14]

$$C(x) = p(C_{census}(x), \lambda_{census} + p(C_{AD}(x), \lambda_{AD}) \qquad (3)$$

with

$$p(c, \lambda) = 1 - exp(-\frac{c}{\lambda}) \qquad (4)$$

to map the cost values to the range $[0, 1]$.

The pixel–wise computed costs are then aggregated within a dynamic local area surrounding the pixel of interest. For this, we utilize the cross–based skeleton method presented in [15]. This combination has proven to allow for fast computation of good quality reconstructions using color images from calibrated cameras only [12].

For this approach we extend the pixel–wise cost function to also take the unwrapped high–frequency patterns projected ($\Phi'_{h,projector}$) and captured ($\Phi'_{h,camera}$) into account. The assumption is that next to the visual difference of the blended color images, the difference of the intensity within the projected patterns allow for a better reconstruction than the color

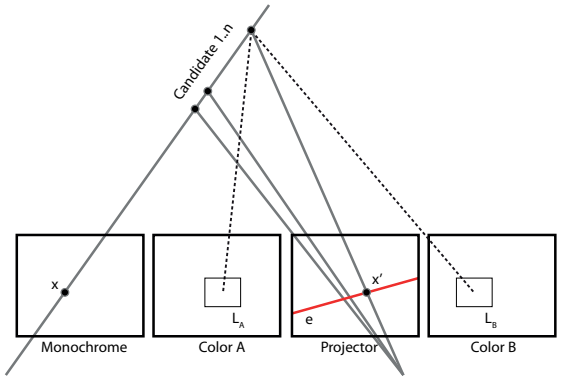

Fig. 5. Candidate search and validation of the SPU approach.

based approach. The new extended pixel–wise cost function $C'(x)$ used for this approach is expressed by:

$$C'(x) = \frac{|\Phi'_{h,projector} - \Phi'_{h,camera}| + C(x)}{2} \quad (5)$$

B. Segmentation Based Phase Unwrapping (SPU)

For the second approach we incorporate a segmentation of the distorted unwrapped high–frequency pattern as well as a color image based search for corresponding segments. If a whole segment can be matched correctly in the undistorted unwrapped high–frequency pattern, unwrapping of the whole segment to the absolute phase according to (1) becomes possible.

The distorted reprojection of the high–frequency pattern captured by a monochromatic high–speed camera is segmented using a connected components segmentation presented in [16]. Knowing the corresponding segment within the high–frequency pattern a correspondence search along the epipolar line e through the undistorted high–frequency pattern given by the geometric calibration is performed. This epipolar line, depending on the hardware setup and pixel–coordinates x of the pixel of interest in the distorted pattern, intersects with up to all thirty–two wraps of the undistorted high–frequency pattern. Whenever the difference of intensity between the distorted and undistorted pattern is below a certain threshold, a correspondence candidate x' has been found.

For all candidates, the three–dimensional coordinates of the candidate are reprojected into the corresponding color cameras. Then, the visual difference within the color images is evaluated using the pixel–wise cost function (5) and the local cross–based aggregation from [15]. The process of candidate selection and validation is shown exemplary for one pixel–cordinate in figure 5. The local aggregations are declared as L_A and L_B in that figure. The minimum visual difference is stored for the pixel of interest along with the segment number corresponding to the segment in the undistorted high–frequency pattern.

Once all pixels of the distorted high–frequency pattern have been processed the final segment assignment for the whole segment in the distorted pattern is achieved by computing a histogram of the stored segment numbers within the whole segment. The segment with the highest support is selected to represent the corresponding segment in the undistorted pattern. Knowing which segment the pixel of interest corresponds to as well as the intensity value in the distorted high–frequency pattern, the pixel correspondence between the distorted and undistorted patterns along the epipolar line is unique within the segment of the undistorted pattern. Thus, we have a good estimate of $\lfloor N\Phi'_l(x) + 0.5 \rfloor$ for equation (1) which, having found to the correct segment, produces an excellent absolute phase Φ in the corresponding segment. The final triangulation is then computed by using the correct intensity match in the undistorted pattern and the pixel–coordinate of interest in the distorted one.

IV. Evaluation

The evaluation of the two presented approaches is done by assuming that the 3D–reconstruction given by the full active six–pattern PMT method [3] can serve as a ground–truth model. For the two approaches, in order to reduce the amount of necessary sinusoidal phase–shifted patterns, the reconstruction quality itself is not assumed to be enhanced. Instead, assuming an accurate estimation of $\lfloor N\Phi'_l(x) + 0.5 \rfloor$ for equation (1), the optimal solution would be an identical reconstruction like computed by the full active method.

The difference between reconstructed surfaces can easily be measured by comparing their depth maps generated by a projection of the final 3D-model into the image plane of one of the cameras. The difference of the depth values reveal the similarity of the reconstructed surface. Thus, we find the mean squared error (MSE) computed in the depth map domain to be a useful quality measure for evaluation. The lower the MSE value, the better the reconstruction aligns to the full active result. The depth values are relative to the defined clipping volume for a cameras field–of–view. Therefore these values lie within the range $[0.0, 1.0]$, defining the nearest visible 3D–coordinate to be 0.0 and the 3D–coordinate at depth 1.0 to be the most distant visible coordinate. Although the non–metric calibration of the cameras does not allow a direct transfer of these values into (milli-)meters, the relation between the values is sufficient for our evaluation.

However, both approaches suffer from the systematic difficulties induced by passive stereo matching. Homogeneous areas, repetitive patterns, reflections and other problems are not reliably matchable within such algorithms and therefore the captured scene has to be feasible in general for passive reconstruction techniques. Otherwise, the passive matching will fail yielding large errors in the final reconstruction.

The scene used for evaluation has not been adapted to be a good candidate for passive matching. Figure 4 shows a color image of the scene as well as the reconstruction result for the whole scene generated by the full active PMT method. Therefore, the results for comparison of the whole captured scene ($MSE_{complete}$), which are given in table I, are currently of low significance for generic scenes. However, to focus on the quality of reconstruction, we also have evaluated the

reconstructions in a smaller area of the scene, the head region of the horse shown in figure 1, where all algorithms can yield adequate results (MSE_{head}). The results of the head region are also shown in table I:

TABLE I
RESULTS OF THE MSE–BASED EVALUATION.

Algorithm	$MSE_{complete}$	MSE_{head}
PCF	0.9091	$0.3826e^{-3}$
SPU	0.2808	$0.4120e^{-5}$

The interpretation of the values given in table I is straightforward and can easily be confirmed visually by the results shown in figure 1. For the complete scene, the PCF approach features errors way too big to be comparable to the full active reconstruction. The scene features homogeneous areas almost everywhere and therefore this bad result is expected. Obviously, the incorporation of the modulo 2π phase into the cost function of a passive reconstruction method is not sufficient for a high–quality reconstruction.

The reconstruction of the whole scene by the SPU approach also suffers a lot from the lack of texture. However, the aggregated matching throughout each segment seems to be way more suitable than adapting the cost function. Much larger areas can be reconstructed with a low MSE value resulting in an at least partially good–looking reconstruction. For scenes featuring a higher degree of texture the ambiguity of the segments correspondencies are expected to be less significant resulting in an even more suitable reconstruction.

The reconstruction of the smaller part of the scene, featuring the head of the horse that can adequately be matched by both approaches, reveals the actual relation between the approaches. While the PCF approach can already produce a result with a low MSE value, the SPU approach shows a very good similarity compared to the full active reconstruction with an accuracy two orders of magnitude better than the PCF reconstruction.

Visually, there is almost no difference between the interior of a correctly matched segment found by the SPU in comparison to the full active reconstruction. This result shows the capability of the SPU approach to achieve a high–quality 3D–reconstruction, at least for well–textured scenes, that is almost as accurate as a full active reconstruction. Also, both approaches compute their respective reconstructions using only the high–frequency triplet of sinusoidal patterns, effectively reducing the acquisition time and there also the possible rate of reconstruction by a factor of two.

V. CONCLUSION AND FURTHER WORK

We have presented two different approaches for a hybrid 3D–reconstruction by incorporating passive reconstruction techniques in an active PMT 3D–scanning system. While the bottom–up PCF approach, described in III-A, can not benefit from the actively induced high–frequency unwrapped phase, the top–down SPU approach, described in III-B, can achieve high–quality reconstructions, at least for scenes suitable for processing by passive reconstruction techniques.

While the acquisition time of just the high–frequency triplet allows for a rate of reconstruction twice as high, the impact of dynamic objects has still to be determined but is expected to be less intense using the segmentation based phase unwrapping approach.

Further work will focus on the borderless segmentation of the high–frequency patterns in order to further complete the resulting 3D–reconstruction. Finally, to approach the ambiguities induced by problematic areas for the passive techniques used, a more sophisticated segment assignment should be derived taking neighboring segments into account.

ACKNOWLEDGEMENT

This research was supported by the Deutsche Forschungsgemeinschaft, DFG, project number Si 673/11-1.

REFERENCES

[1] L. Onural, T. Sikora, and A. Smolic, "An overview of a new european consortium: Integrated three-dimensional television-capture, transmission and display (3dtv)." in *EWIMT*, 2004.

[2] O. Schreer, P. Kauff, and T. Sikora, *3D Videocommunication*. Wiley Online Library, 2005.

[3] K. Ide and T. Sikora, "Real-time active multiview 3d reconstruction," in *International Conference on Computer Vision in Remote Sensing*, Xiamen University, Xiamen, China, Dec. 2012.

[4] T. Weise, B. Leibe, and L. Van Gool, "Fast 3d scanning with automatic motion compensation," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.

[5] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, no. 1-3, pp. 7–42, Apr. 2002.

[6] F. Blais, "Review of 20 years of range sensor development," *Journal of Electronic Imaging*, vol. 13, no. 1, 2004.

[7] E. Stoykova, A. A. Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, and X. Zabulis, "3-d time-varying scene capture technologies - a survey," vol. 17, pp. 1568–1586, 2007.

[8] J. Posdamer and M. Altschuler, "Surface measurement by space–encoded projected beam systems," *Computer graphics and image processing*, vol. 18, no. 1, pp. 1–17, 1982.

[9] J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[10] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[11] C. Everitt, "Projective texture mapping," *White paper, NVidia Corporation*, vol. 4, 2001.

[12] T. Borgmann and T. Sikora, "Image guided cost aggregation for hierarchical depth map fusion," in *International Conference on Computer Vision Theory and Applications (VISAPP)*. INSTICC, Feb. 2013.

[13] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*. IEEE, 1996, pp. 358–363.

[14] X. Mei, X. Sun, M. Zhou, H. Wang, X. Zhang *et al.*, "On building an accurate stereo matching system on graphics hardware," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 467–474.

[15] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 7, pp. 1073–1079, 2009.

[16] K. Wu, E. Otoo, and K. Suzuki, "Optimizing two-pass connected-component labeling algorithms," *Pattern Analysis and Applications*, vol. 12, no. 2, pp. 117–135, 2009.