Object Description Techniques in Real-Life Environments: a Review for the Underwater Scenario

Roberto Di Salvo and Carmelo Pino
Department of Electrical Electronics and Computer Engineering
University of Catania
V.le A. Doria,6 – 95125 Catania, Italy
{roberto.disalvo & carmelo.pino}@dieei.unict.it

***Abstract*—In this paper we present an overview of the most interesting object description techniques which depicts some descriptors that could be used to support higher level components for object recognition, behavior analysis, classification, and clustering. We are focused our attention on those techniques which are suitable to be used in what are known as real-life environments. In particular the underwater environment were taken into account since it shows a lot of difficulties concenring the low quality of the observed scene and the targets themselves (i.e. fish) which are characterized by fast and erratic movements and more degrees of freedom in motion than, for example, people or vehicles in urban environments.**

Keywords-object description; color, texture, motion and contour features; underwater environment; fish descriptors;

I. INTRODUCTION

In the last few years a large number of computer applications have been developed regarding the use of specific target description algorithm on several research fields such as video surveillance for vehicles [1-3], animals [4-6] and humans [7-9] in order to support both detection/tracking tasks and classification/recognition tasks.

While in a standard condition environment (i.e. indoor environment with motionless backgrounds and static light conditions) the recognition of object of interest in the observed scene and the development of description techniques for the identified target may results in a simple task, the characterization of target description techniques in real-life unconstrained settings such as the underwater environment, implicates to handle different effects that usually occur in the observed scenes [24] such as sudden and gradual light changes, unexpected weather conditions variations (e.g. sudden cloudiness, storms and typhoons) which raise a worsening of image contrast, murky water that affect the clarity and cleanness of the water flow due to the drift and the presence of plankton, rapid formation of algae and filth on camera lens and background movements and variations which cause arbitrary changes in the scene. Fig. 1 shows an overview of the considered environments. Also, differently from humans, in the underwater domain the targets (i.e. fishes) show erratic and fast movements (in three dimensions) that lead to frequent changes in size and appearance. For this reasons, the reviewed descriptors have been chosen in order to deal with the peculiarities of fish appearance and motion in their natural habitat making the above task insensitive to variations in the target's position, size, appearance, orientation and scale with respect to the camera.

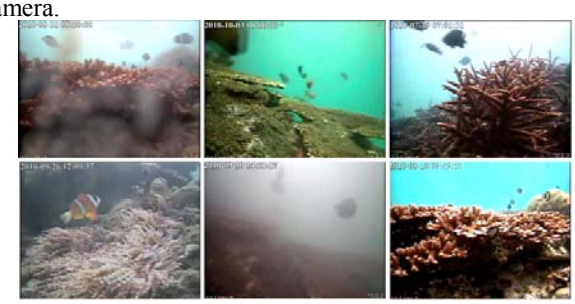

Figure 1. Samples of underwater environments

In the next sections a taxonomy of the descriptor, classified by categories (color features, texture features, motion features and contour features), is proposed and the invariance of these descriptors against different types of changes (e.g. light intensity change) have been investigated for improving both the detection and tracking tasks and the recognition and classification tasks. In particular, for each technique a brief description of what it is, what it is useful for, how it is calculated and what it is invariant to, is given.

II. COLOR FEATURES

A. Joint Histogram

Pass and Zabih in [10] use this method as an alternative to color histograms since it incorporates additional information (local pixel features) without sacrificing the robustness of color histograms. It is a multidimensional histogram created from a set of local pixel features. An entry in a joint histogram counts the number of pixels in the image that are described by a particular combination of feature values. More precisely, given a set of k features (e.g. color, edge, texture, gradient magnitude, rank, etc…) where the i^{th} feature has n_i possible values, we can construct a joint histogram which is a *k-dimensional* vector, such that each entry in the joint histogram contains the number of pixels in an image that are described by a *k-tuple* of feature values. The size of the joint histogram is:

$$n = \prod_{i=1}^{k} n_i \quad (1)$$

that is the number of possible combinations of the values of each feature. Just as a color histogram approximates the

density of pixel color, a joint histogram approximates the joint density of several pixel features.

B. olor Moments

Color moments were treated both in [11] and in [12] as a set of measures that can be used to describe the images features of color. Since the distribution of color in an image can be interpreted as a probability distribution, the moments of that distribution can then be used as features to identify that image based on color. Moments are calculated for each channel in an image, so that it is characterized by 9 moments (3 for each color channel). In fact, most information is concentrated on the low-order moments: the first moment (mean), the second moment (variance) and the third moment (skewness). The three color moments are considered as image features and can be defines as:

Mean: $$E_i = \sum_{N}^{j=1} \frac{1}{N} p_{ij} \quad (2)$$

Variance: $$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{N}^{j=1}(p_{ij} - E_i)^2\right)} \quad (3)$$

Skewness: $$s_i = \sqrt[3]{\left(\frac{1}{N}\sum_{N}^{j=1}(p_{ij} - E_i)^3\right)} \quad (4)$$

A function of the similarity between two image distributions is defined as the sum of the weighted differences between the moments of the two distributions. Formally this is:

$$d_{mom}(H,I) = \\ = \sum_{i=1}^{r} w_{i1}|E_i^1 - E_i^2| + w_{i2}|\sigma_i^1 - \sigma_i^2| + w_{i2}|s_i^1 - s_i^2| \quad (5)$$

where (H, I) are the two image distributions being compared, i is the current channel index (e.g. 1 = H, 2 = S, 3 = V), r is the number of channel (e.g. 3), E_i^1 and E_i^2 are the first moments (mean) of the two image distributions, σ_i^1 and σ_i^2 are the second moments (variance) of the two image distributions, s_i^1 and s_i^2 are the third moments (skewness) of the two image distributions and w_i are the weights for each moments specified by the users.

III. TEXTURE FEATURES

A. Gabor filter

Such a filter described by Kruizinga et al. in [13] is linear and local and is characterized by a preferred orientation and a preferred spatial frequency. It acts as a local band-pass filter with certain optimal joint localization properties in both the spatial domain and the spatial frequency domain. Gabor filters are applied to obtain the G − Maps. A two dimensional Gabor function $g(x, y)$ can be described by the formula (6) as follow:

$$g(x, y, \lambda, \psi, \sigma, \gamma) = e^{-\frac{x^2 + \gamma^2 \cdot y^2}{2 \cdot \sigma^2}} \cdot \cos\left(2\pi \frac{x}{\lambda} + \psi\right) \quad (6)$$

where λ, ψ, σ, γ are respectively the orientation, the scale, the mean and the standard deviation of the considered Gabor filter.

Given an image $I(x, y)$, the Gabor transform is obtained by a convolution between the image I and the function g. It is possible to use 6 scales and 4 orientations, thus obtaining 24 complex images. Then, the mean and the standard deviation of the magnitude of each of these complex images are taken as descriptors.

B. Scale-Invariant Feature Transform (SIFT)

The SIFT standard method proposed by Lowe in [14] is a technique for extracting distinctive invariant features from images, used to perform reliable matching between different views of an object or scene. The gradient of an image is shift-invariant, while under light intensity changes, i.e. a scaling of the intensity channel, the gradient direction and the relative gradient magnitude remain the same. Also, because the SIFT descriptor is normalized, the gradient magnitude changes have no effect on the final descriptor. The features produced by SIFT are invariant to image scale and rotation and provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. Moreover the features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images.

Operatively a keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location. These are weighted by a Gaussian window, indicated by the overlaid circle in the left image above. These samples are then accumulated into orientation histograms summarizing the contents over 4 x 4 sub-regions: the length of each arrow corresponds to the sum of the gradient magnitudes near that direction within the region. The right figure in Fig. 2 shows a 2x2 descriptor array computed from an 8x8 set of samples (left figure).

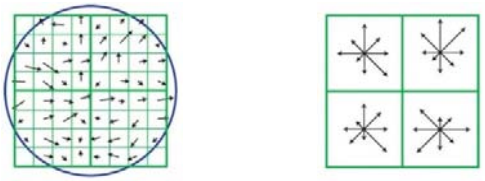

Figure 2. Image gradient (right) and Keypoint descriptor (left) reported from [14]

C. SIFT with Global Context

This approach has been proposed by Mortensen et al. in [15] and consists of a feature descriptor that augments SIFT with a Global Context vector that adds curvilinear shape information from a much larger neighborhood. One of the most evident limit of the classical method is that SIFT typically fail to consider global context to resolve ambiguities that can occur locally when an image has multiple similar regions. With the global context the mismatches can be reduced when multiple local descriptors are similar. In fact, rather than count distinct edge points, with Global Context the maximum curvature at each pixel (x,

y) is computed as the maximum eigenvalue of the Hessian matrix:

$$H(x,y) - \begin{bmatrix} r_{xx} & r_{xy} \\ r_{xy} & r_{yy} \end{bmatrix} \quad (7)$$

where r_{xx} and r_{yy} are the second partials of the image in x and y, respectively, and r_{xy} is the second cross partial.

D. PCA-SIFT

This method proposed in [16] examines the local image descriptor used by SIFT, but instead of using SIFT's weighted histograms, the Principal Components Analysis (PCA) to the normalized gradient patch is applied. To be more precise PCA-SIFT consists of a i) pre-computation of an eigenspace to express the gradient images of local patches, ii) given a patch, compute its local image gradient and iii) project the gradient image vector using the eigenspace to derive a compact feature vector. The latter is significantly smaller than the standard SIFT feature vector, and can be used with the same matching algorithms. For these reasons this method have the advantage that is more distinctive and robust to image deformations than the standard SIFT representation increasing accuracy and providing faster matching.

E. Covariance Matrix

This method proposed by Donoser and Bischof in [17] and Lakmann in [18] is a second order statistics about the correlation of pixel pairs for a set of topological pixel relations. In particular the spatial relations of pixel values can be calculated for the different color components of involved pixels. The spatial relations can be analysed i)inside each color plane, ii) between the color components of pixels in different color planes and iii) for a set of feature built out of each pixel belonging to an object's region. These representations take into account both the spatial and statistical properties, unlike histogram representations (which disregard the structural arrangement of pixels) and appearance models (which ignore statistical properties). The most important characteristic of covariance matrices depends on their independence to variations of colors in image sequences. This problem in color constancy is often caused by separate fluctuations in the brightness of the color channels. So the covariance matrix features could be invariant both for additive and multiplicative noise. The main shortcoming of this approach is the amount of time required for the computation of the covariance matrix; however, approaches based on integral images have been proposed in order to improve the speed.

IV. Motion Features

A. Motion Vector Analysis

It is a representation of the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (e.g. a camera) and the scene [19][20]. The most used methods for motion vectors assessment are typically based on the optic-flow estimation (i.e. Horn–Schunck or Lucas–Kanade approaches), with the advantage that it does not have to find feature point correspondence. The motion vector field (or optical flow in gradient-based approach), is estimated based on the instantaneous change in image intensity. Assuming that the intensity along the motion trajectory is always constant, this estimation often bases on minimizing the optical flow constraint and some local smoothness constraints. This makes optical flow have the property of obtaining the estimated motion field in fine resolution, more close to "true" motion. A drawback of these methods is that the obtained optical flow does not represent the true motion, but only the motion projection on the direction of image gradient. Basically in motion estimation, an object keeps its brightness unchanged in the image sequence, so let the optical flow (motion) at pixel *(x, y)* be *(u(x, y), v(x, y))*, E_x be the derivative of image intensity in x-direction, E_y be the derivative of image intensity in y-direction, and E_t be the derivative of image intensity in t-direction, classical optical flow constraint can be described as shown in (8):

$$u \cdot E_x + v \cdot E_y = -E_t \quad (8)$$

For a particular image point *(x, y)*, the values of u and v are restricted by this linear equation.

V. Contour Features

A. Curvature Scale Space and Curvature Points

The CSS (Curvature Scale Space) method proposed by Ghosh and Pektov in [21] and Spampinato et al. in [4] provides a set of boundary descriptors which can be used to characterize the object's contour even though it is affected by 3D transformation in the observed scene. CSS is executed by iteratively smoothing the curve until the number of points where the curvature is zero (zero crossing points) is equal to zero. The CSS image represents curvature zero crossings during shape evolution in the plane *(u, σ)*, where u is the normalized arc length between consecutive zero crossing points and σ is the width of Gaussian kernel used for shape smoothing. The curvature is defined as the changing rate of curve slope, according to the formula:

$$k(u) = \frac{x(u)\dot{y}(u) - \dot{x}(u)y(u)}{\sqrt{(\dot{x}(u)^2 + y(u)^2)^3}} \quad (9)$$

where u is the curve formed by the computed boundary. To find the CSS image, we iteratively smooth the extracted boundary.

Let *g(u, σ)* be a $1-$ D Gaussian kernel of width σ, then the components of the evolved curve Λ_σ may be represented by *X(u, σ)* and *Y(u, σ)* according to the properties of convolution:

$$X(u,\sigma) = x(u) * g(u,\sigma)$$
$$Y(u,\sigma) = y(u) * g(u,\sigma) \quad (10)$$

where $(*)$ is the convolution function.

The derivatives are:

$$X_u(u,\sigma) = x(u) * g_u(u,\sigma)$$

$$X_{uu}(u,\sigma) = x(u) * g_{uu}(u,\sigma) \quad (11)$$

where $g_u(u, \sigma)$ and $g_{uu}(u, \sigma)$ are, respectively, the first and the second derivative of the gaussian function. The same holds for $Y_u(u, \sigma)$ and $Y_{uu}(u, \sigma)$.

The curvature of the evolved digital curve is:

$$k(u,\sigma) = \frac{X_u(u,\sigma)Y_{uu}(u,\sigma) - X_{uu}(u,\sigma)Y_u(u,\sigma)}{(X_u(u,\sigma)^2 + Y_u(u,\sigma)^2)^{3/2}} \quad (12)$$

As σ increases, the shape of Λ_σ changes. Thus, we have to calculate several times the curvature zero crossing points of Λ_σ during the curve evolution, until when the number of such points will be zero. For each iteration (value of σ) the arc length between consecutive zero crossing points is plotted in the CSS image.

B. *Fourier Descriptor*

These contour features, used by Arbter et al. in [22] and Sener and Unel in [23], are independent form the object's position, orientation, scale and slant since usually fish can be at any position and orientation relative to the camera [ref_artemis]. Operatively the extracted boundary can be expressed as a sequence of coordinates $s(k) = [x(k), y(k)]$, where $x(k)$, $y(k)$ are the coordinates of the points of the boundary. Each pair of coordinated can be considered a complex number, i.e. $s(k) = x(k) + j \bullet y(k)$. The discrete Fourier transform (DFT) is:

$$a(u) = \frac{1}{K}\sum_{k=0}^{K-1} s(k) \cdot e^{-j2\pi uk/K} \quad (13)$$

for $u = 0, 1, 2...K$, where K is the number of points belonging to the identified boundary. The complex coefficients $a(u)$ are the Fourier descriptors and provide a means for representing the boundary of a two-dimensional shape. Since it is not feasible to use all the Fourier descriptors in the classification step due to their high number, in order to describe the frequency variability of the shape we use a histogram of their modulus with 30 values. The histogram of Fourier descriptors is invariant to affine transformation.

VI. CONCLUSIONS

In this paper a survey on object description method for real-life environment have been proposed. The reviewed techniques have been chosen to deal with the challenges of the underwater domain in order to support the users when design high level applications for object recognition, classification and behavior analysis. The considered approaches are ready for production run and could be refined with several techniques (some already available and some to be developed) in order to cover the difficulties related both to the object movements during the observation in their natural habitat (fish in our case study) and the changes in the considered scene triggered by the environmental factors. In the next years, we would expect new work than simply "apply" existing algorithms to the underwater unconstrained research area, in order to demonstrate principles, techniques and real advantages of using the available methods on different outdoor environments.

REFERENCES

[1] Faro A., Giordano D., Spampinato C., "Integrating location tracking, traffic monitoring and semantics in a layered ITS architecture", IET Intelligent Transport Systems 5, 197 (2011).

[2] Faro A., Giordano D., Spampinato C., "Evaluation of the traffic parameters in a metropolitan area by fusing visual perceptions and CNN processing of webcam images", IEEE Transactions on Neural Networks, (6) 1108-1129, 2008

[3] Faro A., Giordano D., Spampinato C., "Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection", IEEE Transactions on Intelligent Transportation Systems, (4) 1398-1412, 2011

[4] C. Spampinato, D. Giordano, R. Di Salvo, J. Chen Burgher, R. B. Fisher, G. Nadarajan, "Automatic fish classification for underwater species behavior understanding", Proceedings of the ACM Int. Workshop Anal. and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS), October 29th, 2010, Florence, Italy.

[5] Spampinato C., Palazzo S., Giordano D., Kavasidis I., Lin F.-P., Lin Y.-T., "Covariance based fish tracking in real-life underwater environment", VISAPP 2012 – Proceedings of the International Conference on Computer Vision Theory and Applications, 409-414, 2012.

[6] C. Spampinato, J. Chen Burger, G. Nadarajan, R. B. Fisher, "Detecting, tracking and counting fish in low quality unconstrained underwater videos, Proceedings of VISAPP 08, January, 22nd – 25th, 2008, Madeira, Portugal.

[7] Faro A., Giordano D., Design memories as evolutionary systems socio-technical architecture and genetics, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 4334-4339, 2003.

[8] Di Salvo R., Faro A., Giordano D., Spampinato C., People flow control using cellular automata and computer vision technologies, Advances in Intelligent and Soft Computing, Volume 159 AISC, Issue1, 2012, Pages 95-104, Future Computer and Control Systems, FCCS, 2012.

[9] A. Faro, D. Giordano, C. Spampinato, "An automated tool for face recognition using visual attention and active shape models analysis", Proceedings of the 28th IEEE EMBS Annual International Conference, Aug 30th - Sept 3rd, 2006, New York City, USA.

[10] G. Pass, R. Zabih, "Comparing images using joint histograms", Multimedia System, 7 (1999), pp. 234–240.

[11] J.-L. Shih, L.-H. Chen, "Colour image retrieval based on primitives of colour moments", Vision, Image and Signal Processing, IEEE Proc., vol.149, no.6, pp. 370- 376, Dec 2002 doi: 10.1049/ip-vis:20020614.

[12] N. Keen with R. Fisher, "Color Moments" Feb. 10, 2005.

[13] P. Kruizinga, N. Petkov, S. E. Grigorescu, "Comparison of texture features based on Gabor filters", Proceedings of the 10th International Conference on Image Analysis and Processing, Venice, Italy, September 27-29, 1999, pp. 142-147.

[14] D. Lowe, "Distinctive image features from scale-invariant keypoints", Int. J. Computer Vision, vol. 60, pp. 91-110, 2004.

[15] E. N. Mortensen, H. Deng, L. Shapiro, "A SIFT Descriptor with Global Context", 2005 IEEE Comput. Society Conference on Comput. Vision and Pattern Recognition (CVPR '05) – Vol. 1.

[16] Y. Ke, R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors", 2004 IEEE Comput. Society Conference on Comput. Vision and Pattern Recognition (CVPR '04) – Vol. 2.

[17] M. Donoser, H. Bischof, "Using covariance matrices for unsupervised texture segmentation" 19th International Conference on Pattern Recognition (ICPR 2008), pp.1-4, 8-11 Dec. 2008.

[18] R. Lakmann, "Textural Features in Multi-Channel Color Images", ACCV 2002, The 5th Asian Conference on Computer Vision, 23-25 Henuary 2002, Melbourne Australia.

[19] P-C. Chung, C.L. Huang, E.L. Chen, "A region-based selective optical flow back-projection for genuine motion vector estimation",

Pattern Recognition, Volume 40, Issue 3, March 2007, Pages 1066-1077, ISSN 0031-3203, 10.1016/j.patcog.2006.06.019.

[20] S. Ali, M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis" Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on , vol., no., pp.1-6, 17-22 June 2007 doi: 10.1109/CVPR.2007.382977.

[21] A. Ghosh, N. Petkov, "Effect of high curvature point deletion on the performance of two contour based shape recognition algorithms", International Journal of Pattern Recognition and Artificial Intelligence Vol. 20, No. 6, pp. 913-924, 2006.

[22] K. Arbter, W.E Snyder, H. Burkhardt, G. Hirzinger, "Application of affine-invariant Fourier descriptors to recognition of 3-D objects" Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.12, no.7, pp.640-647, Jul 1990 doi: 10.1109/34.56206.

[23] S. Sener, M. Unel, "Affine invariant fitting of algebraic curves using Fourier descriptors", Pattern Analysis & Applications Volume 8, Numbers 1-2 (2005), 72-83, DOI: 10.1007/s10044-005-0245-6.

[24] C. Spampinato, S. Palazzo, B. Boom, J. Van Ossenbruggen, I. Kavasidis, R. Di Salvo, F.-P. Lin, D. Giordano, L. Hardman, and R. Fisher. Understanding fish behavior during typhoon events in real-life underwater environments. Multimedia Tools and Applications, pages 1-38, doi: 10.1007/s11042-012-1101-5.