3D Representation for Object Detection and Verification

Luis Villavicencio, Carlos Lopez-Franco, Nancy Arana-Daniel, and Lilibet Lopez-Franco

Computer Science Department, Exact Sciences and Engineering Campus, CUCEI, University of Guadalajara, Av. Revolucion 1500, Col. Olimpica, C.P. 44430
felipe.vive@gmail.com,
{carlos.lopez,nancy.arana,lilibet.lopez}@cucei.udg.mx

Abstract. In this paper we introduce a representation for object verification and a system for object recognition based on local features, invariant moments, silhouette creation and a 'net' reduction for depth information. The results are then compared with some of the most recent approaches for object detection such as local features and orientation histograms. Additionally, we used depth information to create descriptors that can be used for 3D verification of detected objects. Moments are computed from a 3D set of points which are arranged to create a descriptive object model. This information showed to be of matter in the decision whether the object is present within the analyzed image segment, or not.

Keywords: object detection, object verification, visual pattern recognition.

1 Introduction

Object recognition is a challenging task that involves several steps aimed to find a relation between an input image and a set of previously known objects [1]. Recent work has brought techniques that allow detection at real-time. Nevertheless, any method applied to object recognition is likely to find the desired object where there is not, that is, it could throw false positive responses. The false positive rate increases when objects are relatively small or the train data is not rich enough or does not describes the object very accurately. The detection step consist in matching an area or points in the image to a known object model, whilst, a verification step goes further and reinforces the decision taken decreasing the false positive rate. Thus, a verification step refines the object detection checking whether the areas really contain the target objects [2].

As information is added to the detection step it becomes slower and heavier. We analyze the inclusion of 3D information in a verification step after detection with the main objective of reducing the false positive rate. The addition of depth information helps to create descriptors that can be easily used to undertake a verification step successfully. This data is added to a set of keypoints forming a

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2013, LNCS 7914, pp. 54–63, 2013.

3D model from which we compute invariant moments. This approach generates models that are light and descriptive

In this paper, we propose to get a small descriptor based on invariant moments to increases the effectiveness of classification. SURF keypoints [3], and Support vector machine classifiers (SVM) [4] are used. We describe algorithms for contour definition and 3D information reduction into a grid over the object. The paper is organized as follows. In section 2, we describe the tools put together to integrate the recognition system. Section 3 presents the proposed methods. In section 4, we present experimental results. Finally, the conclusions are given in section 5.

2 Background

The use of visual features for object detection is very common and functional. For short, features are points of interest that describe the image and make the correspondence problem able to be solved [5]. Robust features identify objects despite changes in illumination, orientation, translation, scale, noise and distortions. Feature descriptors add information of the neighborhood surrounding the key point. The Speeded-Up Robust Features algorithm (SURF) proposed by Bay et al. [3] finds features using integral images and Haar-like features. SURF features provide robustness and speed compared to similar approaches.

Invariant Moments. Moments provide compact information of a data set. A pattern may be represented by a density distribution function, moments can be obtained for a pattern representing an object, and they can be used to discriminate between objects (or classes)[6]. This technique has been previously used in pattern and object recognition as far as the early 60s [7,6,8,9]. Nevertheless, they were usually applied on measures from an RGB image and the classifiers used were simple. The general two dimensional equation for moments is

$$m_{pq} = \sum_{x} \sum_{y} x^p y^q f(x, y) \tag{1}$$

where $f(x, y)$ is a function of the variables, commonly used functions for images are the gray scale function and histograms, this last one is related to the density distribution function. The order of moments is $(p + q)$ [7,9]. The first order moments can be used to locate the centroid of the set of points.

$$\bar{x} = \frac{m_{10}}{m_{00}}, \quad \bar{y} = \frac{m_{10}}{m_{00}} \tag{2}$$

If we compute moments considering a translation to the centroid, we generate central moments.

$$\mu_{pq} = \sum_{x} \sum_{y} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{3}$$

Central moments, μ_{pq}, can be made invariant under scale dividing them by μ_{00}^{γ}. μ_{00} defines the area so this is a scaling normalization.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}, \quad \gamma = [(p + q)/2] + 1 \tag{4}$$

We use a set of 3D points for the moment functions. If we compute a density function of the form $F_x(X) = \int f(u)\ du$ counting the number of occurrences of every point, the probability value assigned to them will be equal because each point appears only once. Thus $f(x, y)$ is assumed as 1 all the time.

The first four moments are descriptive measures for a distribution. Using moments for visual pattern recognition not on the RGB images but on depth information gives a new perspective on the use of moments, since 3D information describes the distribution of points that conform an object.

3 Object Recognition System

This section is focused on describing the recognition system. First, SURF points are extracted from an input image and matched to known models. Small areas around matches are subject to further inspection. Using features we construct a contour taking points with a sliding orientation window based on the magnitude measure. Next, we proceed to extract depth information within the boundaries of the contour and reduce the number of points that will be used to compute the moments. The reduction is intended to generate a smaller set with rich information and less heavy. This is done adjusting a mesh grid over the object. To do this, a fixed number of points are initialized in coordinates inside the bounding box of the contour then they are migrated to enhance the model. Depth information is attached to the coordinates. Then, we calculate the invariant moments over this reduced 3D set to form a small descriptor. Finally, the descriptor is used for classification using a linear SVM. This process is seen in Fig. 1.

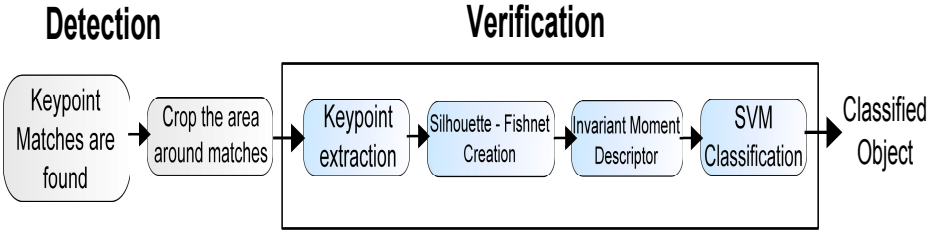

Fig. 1. Steps for the object recognition system

Keypoints. The first step is the extraction of keypoints using the SURF algorithm in an image containing the object, these keypoints are later used to generate a contour. Some other alternatives were tested, including morphological operations and edge or corner detectors but they were not as robust. Feature description resulted the most efficient method for the robustness of points through changes in scene variations. The SURF keypoints used skip descriptor formation which increases speed. Sample SURF keypoints can be seen in Fig. 2.

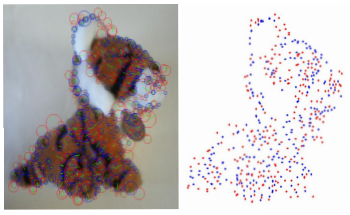

Fig. 2. SURF points extracted for an object

3.1 Silhouette Extraction Algorithm

To create the silhouette from the set of keypoints, we start defining the bounding box of keypoints and find the centroid. It is used to translate and scale the points. Magnitude and orientation are calculated for all points using

$$m(x, y) = \sqrt{(x - x_0)^2 + (y - y_0)^2} \tag{5}$$

$$\theta(x, y) = tan^{-1}\left(\frac{y - y_0}{x - x_0}\right) \tag{6}$$

Where (x_0,y_0) is the centroid. Next, we break the 360 degrees into a number of bins corresponding to the number of points in the contour. For instance, if we want 72 points, bins represent 5 degrees each. A sliding orientation window is used to take points with the highest magnitude for every bin.

Occasionally, some points may be zero because SURF key points were not triggered in certain zones. We scan for blanks and make a linear interpolation to infer missing information using the previous and next known points in the silhouette. Examples of the silhouettes obtained can be seen in Fig. 3.

The number of points in the silhouette has to be carefully chosen. Taking only a few points will create contours with points distant from each other not describing objects correctly. So, small sizes may cause lack of information for descriptors. Very large sizes may cause much sparse information and overuse of interpolation. Since points are selected from SURF features, the contour is limited to the number of keypoints. Taking a large number will produce many blanks.

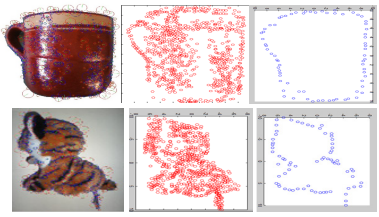

Fig. 3. Sample Silhouettes for two objects

3.2 Fishnet Reduction Algorithm

After silhouette creation, we add depth information. If the area occupied by the object is big, it leads to a large set of points. We perform a reduction of the information adjusting a net over the object. Having depth and RGB images aligned, 3D information is cropped to retain only the area contained inside the bounding box of the silhouette.

First, we define the number of points and generate (x,y) coordinates by sectioning the bounding box evenly. This creates a set of bidimensional points $Grid = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, the value n is a constant that indicates the number of points in a single direction. Besides,

$$\begin{aligned} x_i \in [x_{min}, x_{max}],\ y_i \in [y_{min}, y_{max}] \\ x_1 = x_{min}, \qquad y_1 = y_{min} \\ x_n = x_{max}, \qquad y_n = y_{max} \\ x_{i+1} = x_i + x_{inc},\ y_{i+1} = y_i + y_{inc} \\ x_{inc} = \frac{x_{max} - x_{min}}{n}, y_{inc} = \frac{y_{max} - y_{min}}{n} \end{aligned} \tag{7}$$

minimum and maximum values are defined by the contour bounding box. The result of this process can be seen in Fig. 4 (up).

Since we take the bounding box measures some points might lie outside the contour. The second step is to check that points lie inside the contour. Points are valid if there exist both higher and lower values in points from the silhouette on the vicinity of x and y dimensions. Then, for each non-valid point we take two valid points and move the invalid one between them, this can be seen as a biased migration. The resulting net can be seen in Fig. 4 (down).

For the third step, we add the z coordinate (depth) to the valid (x, y) points. Since depth and RGB images are aligned we append the depth information

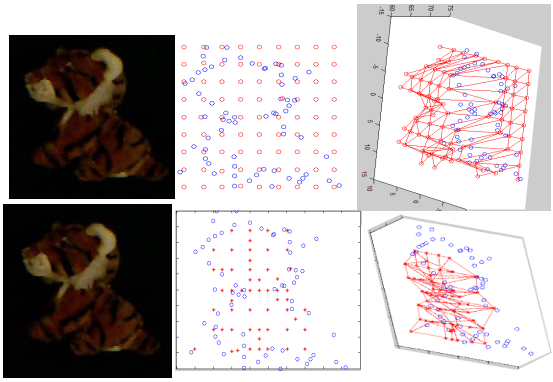

Fig. 4. Net creation. Up, Some points lie outside the contours. Down, points outside the silhouette are migrated.

located at or closest to (x,y). We check that z is a non-zero quantity. When we find null depth points, we look for the next closest point to the coordinates. At the end we have a 2D set of points for the silhouette and a set of 3D points arranged in a net over the object.

3.3 Descriptor Formation

We use the four first invariant moments of each x, y, and z to create the descriptor. The first moment is the mean of the distribution. The second moment is the variance of the set which tells their spread. The third moment determines skewness which measures the symmetry on the shape. The fourth moment defines kurtosis which measures how flat or peaked a distribution is. This results in a 12 elements vector with rich information about the structure of the distribution. After the moments are calculated the vector is used as an input to a linear SVM which classifies the object.

4 Experimental Results

A data set was created composed by 9 objects (cups, totems, tigers, tennis, console controls, globes, shoes, hair dryers and irons) with around 50 images each, see Fig. 5. The images included changes in the scene conditions. For every object 15 images were used to train the SVMs and 25 images for the test stage. The silhouettes created were formed by 72 points, using bins of 5 degrees. On 3D experiments, the fishnet consisted of an 81 point (9x9) mesh. The classification tests were made using cropped, segmented images that included one object and had a discriminative background.

4.1 2D Classification Results

The first experiment consisted in the comparison of 2D contours for invariant moments and Histograms of Oriented Gradients (HOG) descriptors [10] using

Fig. 5. Objects conforming the data set

SVMs classifiers for both. The descriptor was constructed using the information from the first four moments and SURF keypoints (x, y, gradient orientation and scale) generating a 296 element vector. The test consisted of binary classification, with one object being discriminated from another, 1 vs 1, see table 1. In the second test, the long descriptor was compared to a small one with only invariant moments. This test consisted of binary classification using the 4 objects with an object being discriminated from the rest, 1 vs all, see table 2. The results are summarized in result tables. Each row indicates, the percentage of: (a) correct recognitions (CR), (b) false positives (FP), a different object is classified as the target, and (c) false negatives (FN), the target object is classified as a different one.

Table 1. 2D 1 vs 1 Test

1 vs 1 Classification Percentages						
Test	SURF-Moment Desc.			HOG descriptor		
Name	CR	FN	FP	CR	FN	FP
Cups and Tigers	73%	9%	18%	76%	10%	14%
Cups and Tennis	97%	3%	0%	94%	3%	3%
Cups and Totems	92%	8%	0%	97%	0%	3%
Tennis and Tigers	88%	4%	8%	96%	0%	4%
Tennis and Totems	100%	0%	0%	95%	5%	0%
Tigers and Totems	97%	3%	0%	73%	9%	18%

Table 2. 2D 1 vs All Test

1 vs all Classification Percentages						
Test	SURF-Moment Desc.			Moment descriptor		
Name	CR	FN	FP	CR	FN	FP
Cups	88%	10%	2%	84%	6%	10%
Tigers	88%	10%	2%	86%	9%	5%
Tennis	84%	0%	16%	86%	0%	14%
Totems	80%	10%	14%	86%	4%	10%

For the first test, which compares contours and HOG, both SURF-moment and HOG descriptors achieved similar results which shows that moments are a good measure for classification. The second test analyzes classification using moments alone. We can see that SURF information inclusion does not cause a great difference in results. Since Moment-only descriptors throw similar results, we could spare the addition of SURF information. A final test was made in multi-class classification where objects were classified all at once comparing again SURF-moment and HOG descriptors. In this case HOG descriptors made a correct classification of 56% of the objects while the Silhouette-Fishnet-Moment approach obtained a 65% of correct classifications.

4.2 3D Classification and Detection Results

The data set was extended to include depth information. RGB and depth images were aligned and the whole silhouette-fishnet-moment approach was applied. Classification was performed using a very long descriptor (SURF and moments information) and then a simple 12 element vector that only included the first four moments for x, y ans z. In the first experiment, table 3, we compared the two descriptors in order to determine the effect of 3D data inclusion on the computation of moments. Binary classification with two objects was performed. The second experiment, table 4, shows a comparative between the reduced set and the complete 3D set for moment generation, to determine how the reduction affected the pattern. Binary classification with 5 objects was performed for each test, 1 vs all.

Table 3. Results for the 3-D 1 vs 1 Tests

1 vs 1 Classification Percentage						
Test	SURF-Moment Desc.			Moment descriptor		
Name	CR	FN	FP	CR	FN	FP
Hair Dryer and Tigers	85%	5%	10%	100%	0%	0%
Iron and Cups	88%	8%	4%	98%	2%	0%
Shoes and Cups	82%	8%	10%	95%	0%	5%
Iron and Tigers	80%	10%	10%	92%	4%	4%
Globe and Tigers	78%	12%	10%	89%	11%	10%

Table 4. Results for the 3D 1 vs all Tests

1 vs all Classification Percentage						
Test	Reduced Set			Full set		
Name	CR	FN	FP	CR	FN	FP
Hair Dryer	90%	6%	4%	92%	2%	6%
Iron	90%	8%	2%	78%	12%	10%
Tigers	92%	4%	4%	85%	10%	5%
Cups	100%	0%	0%	80%	10%	10%
Shoes	90%	6%	4%	82%	10%	8%

In the first test, the extra information did not increase efficient classification but only made the descriptor heavier. The classification performed by the descriptors using only moments was better for all tests, 3D information turned out to enhance the classification process greatly. In the second test, we can see that using the full set for descriptors did not end in better results. Using the reduced set also reduces time needed to compute moments and improves classification. Finally, a multi-class 3D classification test was performed using the 12 element descriptor and a 70% of correct classification was attained, which shows that depth information enriches the descriptor.

When classification tests were finished, we created a system for object recognition. Using a set of images taken at not preprocessed scenes. The recognition system described in the previous section was implemented. Detection was performed by SURF feature matching and for verification we used the silhouette-fishnet-moment approach. The experiments detected only a class of object at once. An example of the process is seen in Fig. 6. Results are shown in table 5, where we can see that the verification step helps to reduce false positives and enhances detection.

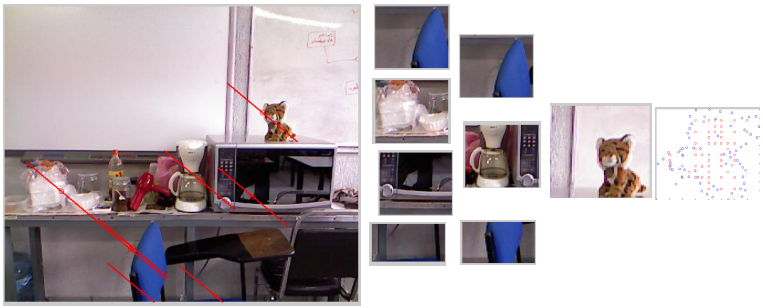

Fig. 6. Recognition system steps exemplification

Table 5. Results for the Object Recognition

Recognition system results				
Object	Matches Found	CR	FN	FP
Hair Dryer	51	94%	2%	4%
Iron	38	89%	6%	5%
Tigers	57	98%	0%	2%
Cups	41	86%	7%	7%
Shoes	36	86%	6%	8%

5 Conclusions

In this paper we presented an object recognition system based on an invariant moment descriptor that includes depth information. Due to the addition of 3D information, invariant moments make small and robust descriptors, outperforming larger descriptors. The inclusion of depth information improves object classification tasks in a huge way. Finally, when a verification step is performed after detection, an important reduction on false positives is achieved. This is very important in many computer vision applications

Acknowledgments. The authors would like to thank CONACYT CB-156567, CB-106838 and University of Guadalajara.

References

1. Jain, R., Kasturi, R., Schunck, B.G.: Machine Vision. McGraw-Hill, Inc. (1995)
2. Baerveldt, A.J.: A vision system for object verification and localization based on local features. Robotics and Autonomous Systems 34(2-3), 83–92 (2001)
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
5. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An Invitation to 3-D Vision: From Images to Geometric Models. Springer (2003)
6. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8(2), 179–187 (1962)
7. Belkasim, S., Shridhar, M., Ahmadi, M.: Pattern recognition with moment invariants: A comparative study and new results. Pattern Recognition 24(12), 1117–1138 (1991)
8. Mercimek, M., Gulez, K., Mumcu, T.V.: Real object recognition using moment invariants. Sadhana 30, 765–775 (2005)
9. Rizon, M., Yazid, H., Saad, P., Yeon, A., Shakaff, M., Saad, A.R.M., Mamat, R., Yacoob, S., Desa, H., Karthigayan, M.: Object detection using geometric invariant moment. American Journal of Applied Sciences 2, 1876–1878 (2006)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893. IEEE Computer Society, Washington, DC (2005)