Unknown objects drawing using image retrieval

Dibujo de objetos desconocidos utilizando recuperación de imágenes

Karina Ruby Perez Daniel[1], Enrique Escamilla Hernandez[1], Takayuki Nagai[2], Mariko Nakano Miyatake[1]*

[1]IPN ESIME Culhuacan. Av. Santa Ana 1000, Col. San Fransciso Culhuacan, 04430, Mexico-city, Mexico.

[2]Universidad de Electro Comunicaciones, 1-5-1 Chofugaoka, Chofu-shi, Tokio 182-8585, Japón.

(Recibido el 1 de mayo de 2011. Aceptado el 10 de noviembre de 2011)

Abstract

In this paper an unknown objects drawing model is proposed. Here we describe a technique that let us build a visual model of a word through images retrieved from Internet, enabling to learn any object at any time. This process is done without any prior knowledge of the objects appearance. However this information must be filtered in order to get the most meaningful image according to the keyword and allowing making the visual relation between words and images as much unsupervised as it could be possible like humans understanding. For this purpose Pyramid of Histogram of Oriented Gradients (PHOG) feature extraction, K-means clustering and color segmentation is done, and then the final image is drawn as an application of the learning process. The proposed model is implemented in a robot platform and some experiments are carried out to evaluate the accuracy of this algorithm.

----- *Keywords:* Learning Objects, PHOG, segmentation, K means, tag object

Resumen

En este trabajo se presenta un modelo para el dibujo de objetos desconocidos. Aquí se describe una técnica que permite construir una idea visual del significado de una palabra, mediante imágenes recuperadas a través de internet, haciendo posible el aprendizaje de cualquier palabra en cualquier momento. Este proceso es realizado sin previo entrenamiento ni conocimiento

* Autor de correspondencia: telefax: + 52 + 1 + 50 + 5 656 20 28, correo electrónico: krperezd@hotmail.com (K. Perez)

de la apariencia de los objetos. Sin embargo ésta información debe ser filtrada con el fin de construir el modelo visual a partir de las imágenes más representativas de la palabra de entrada y así poder generar de manera automática, la relación entre palabras e imágenes de forma no supervisada, tal y como lo hacemos los humanos. Para éste propósito se realiza una extracción de características utilizando el descriptor Pirámide de Histograma de Orientación de Gradientes (PHOG), estas características son agrupadas mediante el algoritmo K medias, seguido de una segmentación de color y una segunda extracción de características. De esta manera la imagen final obtenida es dibujada, representando así la comprensión que se tubo de la palabra de entrada. El modelo propuesto fue implementado en un robot y algunos experimentos son presentados para evaluar la precisión del algoritmo.

----- *Palabras clave:* Aprendizaje de Objetos, PHOG, segmentación, K medias, Objeto Etiquetado

Introduction

In these days the robotic is immersed at most aspects of our life and robots are used to emulate human intelligence not only to make our life easier but also for entertainment purposes [1-3]. So as to achieve this aim, these systems have to get the knowledge from approaches in database training or even from themselves.

Nowadays there are several models for learning objects, most of them are based on training database to make an association network [4] or in image annotations [4,5]. However those algorithms combine image features, textual information or training based on labels, which reminds a semantic problem. There are models which are based just on image features like [6], nevertheless most of those algorithms use manually gathered images, so in some way they are limited from the amount of objects categories stored in their databases or by the accuracy of their textual annotations. Moreover [7] showed a model that gathers images from internet for leaning objects purposes but this refers to build a model from visual words through local features.

The proposed approach in this work is to get learning objects, without any prior knowledge of their appearance, i.e. the images are gathered from the vastest available source, such as the Internet.

For humans, understanding the meaning of the objects is based on images, in this way when we hear the word “Table”, an image of table comes to our minds. Once an image is obtained from the keyword, robot can paint what it learns. However, painting task is not so interesting, but if this goal could be done in an interactive way and it becomes more attractive and the most important, because the robot can learn the appearance of the requested object. The problem is how you draw any object specified by the user. Is not possible for the robot to have prior knowledge of all the objects, so we have to consider how to draw any unknown object (this concept refers to the object is unknown for the system even when for humans is obvious its appearance). This task could be regarded as an unknown object learning problems. In the field of computer vision in recent years, it has been shown to be effective and is actively acquiring the image database using an object concept [8-13].

Thus, taking advantage of “Google Image Search”, retrieved images are downloaded according to the keyword specified by the user. However, from Internet a myriad of images are gathered from the same query, and often contain some images totally unrelated with keyword.

In order to filter the image database obtained from internet and achieve the acquisition of object concepts, all images are clustered using K-means

method based on the similarities of their PHOG (Pyramid of Histogram of Oriented Gradients) descriptors [13]. Traditionally the acquisition of object concept is done through clustering methods [14, 15]. There are several clustering algorithms and several extensions of K-means for some specific types of images or scenes, however, the original K-means is still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, speed and empirical success are the main reasons for its popularity [15]. While, if you apply the concept acquisition algorithm and a probabilistic model for complex instead of clustering method, this becomes a problem for computation time. In this application the goal is to prove that drawing depicts the visual concept of the keyword. Hence, once that all images are described using PHOG, all those images are clustered using K-means where, K is the number of clusters.

The robot can move the pen to draw in accordance with the coordinates of the point. However, it is an important factor in order to draw the edges. There is also a question of efficiency and natural movements like humans at drawing time. Therefore, clustering the edges, must consider the union of near edges and bound image before long drawing time. In this paper, we proposed to implement this algorithm into a real robot for drawing purpose as a playmate.

Traditionally, attempts to draw a picture by a robot [16-18], have been done from a given picture or basically intended to draw the image obtained by the camera of the robot. Talking about human face drawing [18] by a robot is also limited by a prior given picture. These studies have considered a drawing for fun or as an intend of developing human abilities in robots, while according to our point of view; we are trying to draw any object even if the object is unknown to the robot.

Unknown objects drawing imply learning of the objects. Even the proposed scheme was designed for drawing purpose, it can be used or employed to realize other tasks, such as retrieval engines of similar images, object recognition and localization in a given scene.

System summary

Figure 1 shows the outline of the proposed drawing system.

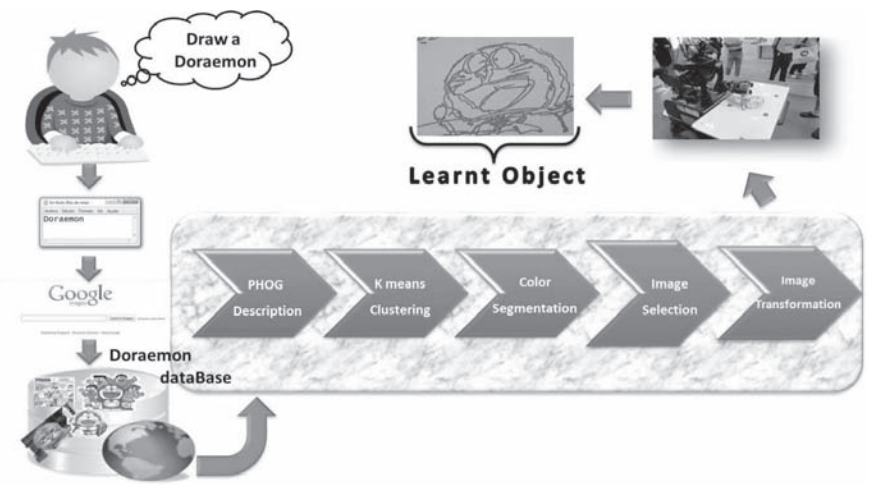

Figure 1 System overview

Where Google Image Search is used as the image learning source and the final drawing depicts the image of the keyword. This learning approach can be divided into 3 phases:

- Image database acquisition
- Image clustering using PHOG and K-means
- Object Segmentation

Learning objects phase

Image database acquisition

Even for Intelligent systems, being able to learn any object by itself is not easy task. Unsupervised learning regards to learn just from similarities between the elements of the database without any prior training. This process depends largely on how big is the image database source and also from the accuracy of retrieved images according to the keyword. Starting from that premise, the robot gets the image database by redirecting the query specified by user to the biggest image source available.

Taking advantage of "Google Image Search" the image database is gathered from there by downloading the first 64 images. Because the proposed unsupervised learning object process depends, largely, on the concordance between the keyword and the most of images from the database, it is necessary to eliminate junk images which are irrelevant to the given queries. From [10] it is well known that Google image search retrieves some unrelated images, however, nowadays the accuracy of retrieved images by Google has improved and the average of unrelated images has been decremented, i.e. most of images retrieved could depicts the keyword. In order to discard the junk images, PHOG, K-means and object segmentation is done.

Image clustering using PHOG and K-means

In this paper, we use K-means clustering [14] for image filtering. To find the most common parameters of the tag object asked at first, each image is converted into feature descriptor using PHOG method [13]. PHOG is a spatial shape distribution of edges descriptor applied into image classification recently [10 - 12] it is possible to obtain the description of the global shape of each object as a vector representation.

PHOG depicts an image descriptor based on pyramid representation which consists of several levels and each of them consists of several cells. During the first level a Histogram of Oriented Gradients (HOG) is applied into the original image, while during the subsequent levels, the image is split into four non overlapped windows and a HOG is done for each new cell. Once that a HOG vector is obtained from each cell and from each level, the final PHOG representation is the concatenation of each single HOG vector.

This method extracts the edge contours using a Canny edge detector and splitting the image into cells, getting one HOG for each grid. Once that all HOG's are calculated then all those vectors are concatenated into only one, called PHOG vector. The orientation gradients are calculated using HOG method, which process the edge contours into the original image using 3 x 3 Sobel mask without Gaussian smoothing. The distribution of each edge provides a specific weight according to the magnitude and is assignment to neighboring bins. The final PHOG is a concatenation of all HOG vectors, and this vector introduces the spatial information of the image. PHOG gives the ability of either matching objects by global shape (edge) or detecting spatial local features and learning correspondence [12, 13].

Thus PHOG is a representation of a component in the outline of the image histogram, which depicts the global characteristics of objects in the image. In addition, a combination of pyramid distribution is done. There is a feature that maintains a degree of spatial information. Here, the gradient or direction component is quantized into n histogram bins.

Figure 2 shows some images and their own PHOG representation when the pyramid level is $L=0$.

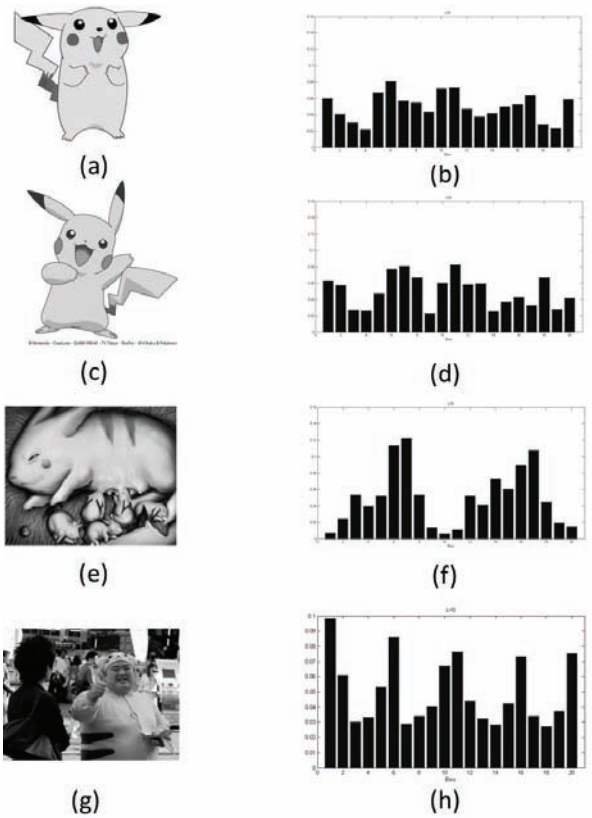

Figure 2 Example of retrieved images and their PHOG representation (level *L=0*)

Here the final PHOG is a 20-dimensional vector, according to the equation number 1:

$$PHOG\,vector\,size = N = \sum_{l \in L} 4^{l} \qquad (1)$$

Where N is the number of bins, l is the current level and L is the total number of levels.

As we can see in figure 2, similar images in shape produce quite similar PHOG's level 0 distribution. For instance the PHOG vector shown in figure 2b is similar to figure 2d, and by the other hand the representation of figure 2f and 2h are so different, because the images that they depicts are also so different.

Thus in order to get the most common group of images according to global shape of the requested object, the pyramid level used is *L=0*.

By the other hand when pyramid level is *L=2*, a more detailed representation, so using this level is possible to distinguish differences between intraclass images. Figure 3 shows four images with similar shape appearance; however their PHOG level 2 representation is different.

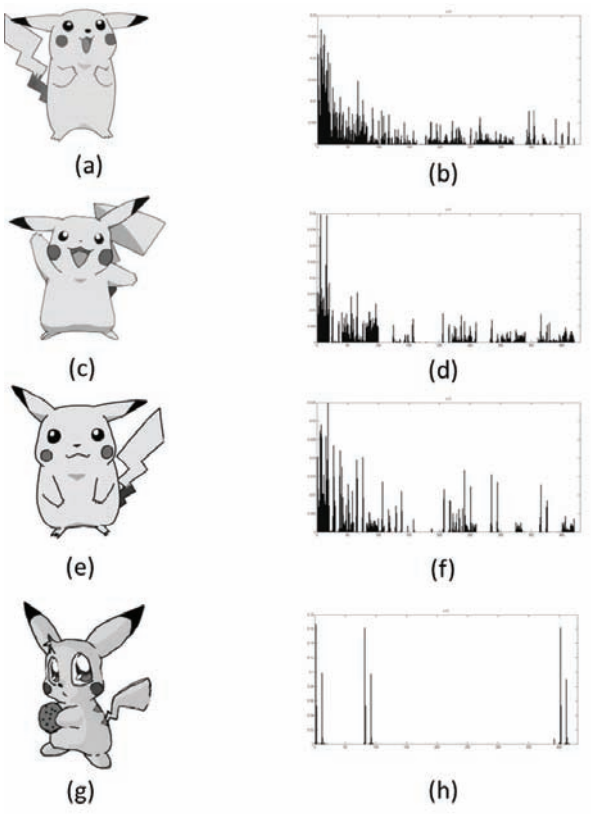

Figure 3 Images and their PHOG vector (level *L=2*)

Once that final PHOG vector is calculated for all the stored images in the database, the next step is to cluster those vectors according to their own similarities. K-means method requires the number of clusters as an input, and then it choses *K* points randomly as cluster centers and assigns each instance to its closest cluster center using Euclidean distance, where *K* is the number of clusters. The next step is to calculate the centroid (mean) for each cluster and use it as a new cluster center, until the cluster center continues being the same.

Worth noting that after the preliminary test were found that the performance does not improve

much even more when $K>3$, so is selected $K=3$ and the bulkier set of images has been the tag object generally, there are others set that contain the object but usually those sets also contains another objects. For that reason the image database is grouped into three sets of images taking in account the similarities between their PHOG vectors, using K-means algorithm and selecting larger clusters, for further processing and others are discarded.

However, what happen when the user query implies an ambiguous concept, i.e. there are some images who are close in semantic but far in appearance. For instance, from the "mouse" query were downloaded 64 images, of which 52 images depicted a computer mouse, 10 out of 64 were a mouse animal and only 2 images were a Mickey mouse. In this case, the most of retrieved images regards to a "computer mouse", i.e. the learning object is based on computer mouse to represent the most.

In this case, the PHOG vector obtained from those images far in appearance is so different, as is shown in figure 4. K-means, clusters into only one group, computer mouse PHOG vector for been similar and for been the higher volume group those images are keep it for a further process, while others are discarded.

Object segmentation

Some images are chosen by the clustering of above, however some of them often have a complex background remains something other than the object. Therefore, segmentation is done in order to separate the object from the background using color information to keep images where the tag object depicts the main shape into the image and the others are discarded. Many of the tag objects are located in an area around the center of the image. Taking into account this premise, we suppose that the region of interest (ROI) is located into the center of the image. Thus, ROI is taken as a reference set of color information. Segmentation was performed in HSV color space and the Euclidean distance. However, if the area of the extracted region (object region) is less than the pre-established threshold, then this image is consider as a likely no object of interest, so is discarded. The threshold value is established by a heuristic way, retrieving more than 1000 images using 50 keywords indicating cartoon's characters, animals and several objects. Figure 5 shows an example of results of segmentation when the keyword given as input was "Pikachu". And the discarded images are labeled with cross.

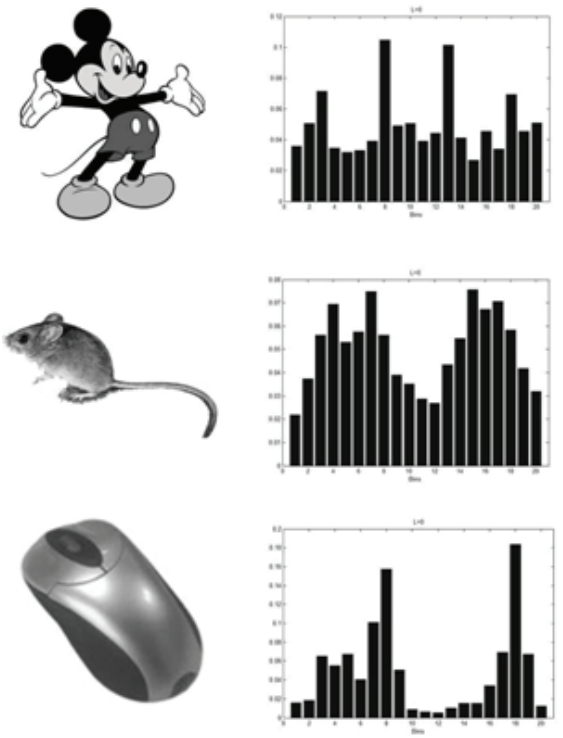

Figure 4 Example of retrieved images by the "mouse" query and their PHOG representation ($L=0$)

After the segmentation of images corresponding to the first clustering, the second clustering using k means and PHOG method is done, so as to obtain the final tag object's group. Unlike the first clustering, in the second one was proposed $K=2$ by setting. Thus, the image database is filtered and the final set of images is landed. The learning process is finished by selecting an image from this final set of images. The final image is the centroid of the dominant cluster found during the second k means process this image is labeled as the "Final Tag Object".

Figure 5 Segmentation results

Figure 6 shows an example of set of images obtained using this method and the Final Tag Object is ticked.

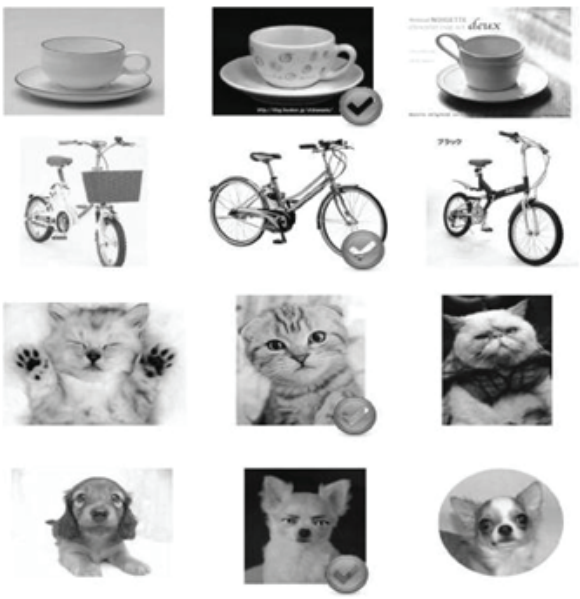

Figure 6 Example of the images obtained

Once the Final Tag Object is achieved, we have the relationship between the initial keyword and its image. This image is applied into a "Drawing Robot".

Drawing robot

Robot Platform

The robot platform is "Digoro". This robot has 6 degrees of freedom (DOF) dual arm, a 2-DOF neck and a waist of one degree of freedom. Lower body and its truck are based on electric wheelchair for the online SLAM (Simultaneous Localization and Mapping) is able to move about freely. The infrared Charge Coupled Device (CCD) camera was calibrated in order to recognize a canvas for drawing. Digoro is equipped with 5 PC-board units, thus it has been coordinated through TCP/IP interface.

All calculation process of the Robot is done by PC to ensure full autonomy, not detrimental to the basic operation of wireless communication in a precarious position.

Image conversion for rendering

So as to develop the drawing stage, edge extraction is done using Canny edge detector. To convert this edge information into the track of the robot arm, at first is necessary to replace the time serial information. So we extract by labeling of edges, find the edge binding connection it requires to convert the time serial information of the edge pixel position into the chain code for each connection. In order to binding the edges of the image, edge detection is achieved by scanning for raster images, when you draw in detection order, means that the point of drawing spatially is close to the edge, the drawing time is short so the hand movements of the arm are efficient, but has the advantage of drawing rather short time to some extent for that reason the drawing process appears little unnatural.

Then, after a long edge binding sort, draw the longest edge at first and then drawing the longest edge until the end.

Thus it is possible to draw with efficient arm movement. In fact, here is the comparison between the 3 ways to draw the order of the edges, which are drawing by length, drawing lines haphazardly and the proposed method so that is drawing by the order of the connected edges. Figure 7 (up) shows the results of the accuracy. This is how the robot has to draw. However even the robot always draws, the minimum arm movements are required and image size is normalized by the width of the drawing square found. Figure 7 (down) shows the drawing ways.

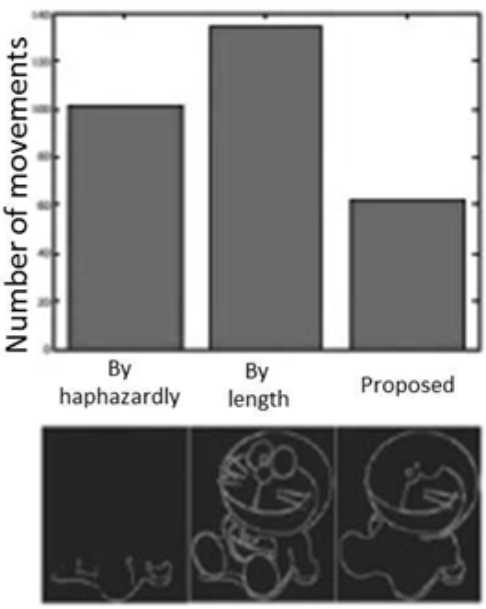

Figure 7 Efficiency of drawing method

Image conversion for rendering

Information obtained by the above algorithm is serial information of the pixel position coordinates of the image. This coordinate system needs to be converted into coordinates to move the robot arm. In this paper, we assume that a drawing process is done on a flat surface, with 4 markers recognized by the three-dimensional color camera attached to the first corner on the canvas, to get the 3D position. Pixel position series can be converted into 3-D position coordinates of the robot, by mapping the four corners of the image pixel position of this three-dimensional location.

The robot gets the information by serial way and the drawing process is done to control the arm position. In this case, the robot grasps the pen with the hand and draws according to the given coordinates. With respect at the hand mechanism, it's difficult to draw gripping the pen as usual way, so the robot draws up the pen of the thickness which adjusts to the robot's hand. In addition, in the robot's hand is placed a pen to draw and the pen moves with the hand in order to develop a better drawing. However, the work of painting robots for home uses is considered important that the robot must be able to draw grasping a normal pen.

Experimental results

From the first stage is obtained a set of images which share similar PHOG features. During segmentation both images where the tag object occupy a small area and images with complex background are discarded, and finally from the second clustering are kept images with very similar PHOG (level 2) features as is shown in figure 8. In our research the descriptor is an oriented histogram and it was quantized into 20 orientation bins in the range of [0, 180] gradients computed on the output of a Canny each detector, 2 levels and 20 bins, thus the descriptor obtained is a 420-vector for this stage according to equation 1.

Once that final data set is achieved, an image of this set is picked up. This image is selected in relation to the centroid found during the last k means clustering process and it is taken as the "Learn Object" and it should be appropriate according to the initial keyword (For instance, figure 9).

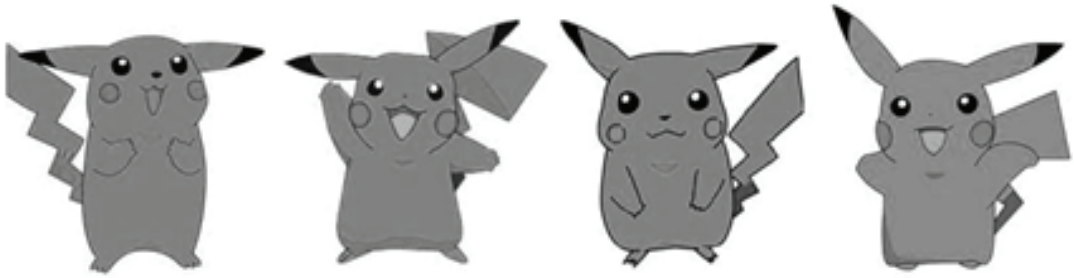

Figure 8 Final data set

Figure 9 Learn Object

The next subsection shows the both results learning and drawing process.

Unknown object learning

For the experiment of an unknown object learning (image selection), was evaluated the learning of different objects from several categories under "Ground Truth" test. This evaluation was done for 30 randomly keywords, such as 10 general objects like planes, cars, teapots, etc., from 10 animals, and from 10 animated characters. We select these images and we check if the image obtained really concern to the asked image. Thus a subject who does not know the edge picture which finally is obtained should decide if it concerns to the initial keyword. And if the answer to what a painting was right, you could not as incorrect demanded accuracy rate. Table 1 shows the results of this evaluation.

After the evaluation of other categories, the next results shown in table 2 were yielded.

Table 1 Results of learning process evaluation

Class name	*Wrong classified objects (final data)*	*Success of the learn object*
Teapot	0	Yes
Car	0	Yes
Airplane	0	Yes
Bicycle	0	Yes
Table	1	Yes
Chair	1	Yes
Flower	3	No
Cup	0	Yes
Pencil	1	Yes
Notebook	0	Yes
Rate of success of this category		90%

Table 2 Results per category

Category Name	*Success of the Learn Object*
Cartoon's Characters	80%
Animals	100%
Varied Objects	90%

The above tables depict the level of understanding success. However after several evaluations we conclude that for more specific keywords,

the accuracy of the learn object improves. For instance in Table 1, the "Flower" class depicts the worst case, however if this word become more specific, the results also improves, as is shown in table 3. Note that now "Flower" is a category with several classes unlike to the table 1, where "Flower" is a class.

However, understand that it may contain non-ASCII characters and background objects. When, on the drawn picture we are able to distinguish the object, it is considered as a successful learning. Nevertheless when the tag object is located into a complex background (texture environment), is difficult to distinguish of the object. Actually correct answer ratio, with the general object was 90%, with the character of cartoons 80% and finally with animal class, became 100%. Worth noting that if the object name is very specific the rate of success improves around 10%. Figure 10, shows some results our learning object method.

Table 3 Results of Flower Category

Class Name	*Success of the Learn Object*
Rose	Yes
Tulip	Yes
Alcatraz Flower White	Yes

Figure 10 Some results of learning process

Figure 11 shows some drawn images and their keyword.

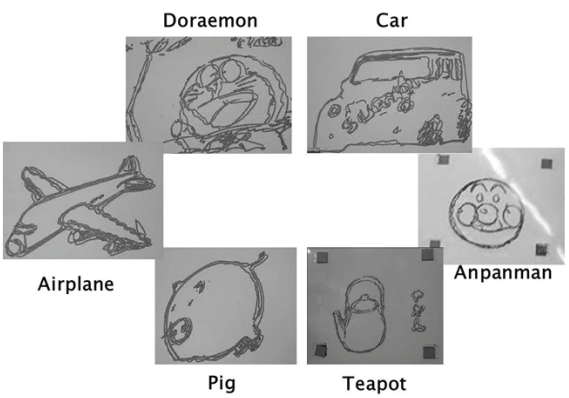

Figure 11 Drawn Images

Conclusion and future work

In this paper, the picture drawn by the robot as one function of the playmate robot was examined. It is possible for the robot to draw images, even when the appointed keyword represents an unknown object for it, so as to learn the real appearance of this object some pictures are retrieved from Internet, and a filtering model was developed based on PHOG features.

PHOG level 0 features provide the main features patterns to get the principal image data set. By the other hand due PHOG level 2 is possible to find the final image data. Using this technique a

more detailed vector representation is achieved and make us easier to find intraclass differences and finally due K-means clustering those vectors can be classified according to their intraclass similarities. The images from the final data set, share similar appearance features and one of them is picked up and taken as the image of the initial keyword.

Thus we proof that following the proposed approach it is possible to learn the appearance of any object, i.e. make the association between words and their image like humans done, without previous image database stored. In this way the drawn image depicts the robot understanding of the keyword.

However, the current picture selected, not contain criteria to make easier the drawing task and sometimes this image is a very detailed picture.

As a future work and in order to improve the scope of this project, we propose to add speech recognition or phoneme recognition. By the other hand this method also could be applied into other kind of tasks for example in robot navigation systems and even into image retrieval engines based on similarities.

Finally, our aim is to achieve a robot and a playmate for the children and the elderly in combination with various other games.

References

1. K. Dautenhahn, I. Werry, J. Rae, P. Dickerson, P. Stribling, B. Ogden, "Robotic Playmates: Analysing Interactive Competencies of Children with Autism Playing with a Mobile Robot". *Socially Intelligent Agents - Creating Relationships with Computers and Robots*. K. Dautenhahn, A. Bond, L. Canamero, B. Edmonds (editors). Ed. Kluwer Academic Publishers. Boston. 2002. pp.117-124.

2. A. M. Howard, H.W. Park, C. C. Kemp. *Extracting Play Primitives for a Robot Playmate by Sequencing Low-Level Motion Behaviors*. Proc. of IEEE Int. Symp. on Robot and Human Interactive Communication. Munich, Germany. August 2008. pp.360-365.

3. A. J. B. Trevor, H. W. Park, A. M. Howard, C. C. Kemp. *Playing with Toys: Towards Autonomous Robot Manipulation for Therapeutic Play.* Proc. of IEEE Int. Conf. on Robotics and Automation. Kobe, Japan, May 2009. pp. 2139-2145.

4. H. Fu, Z. Chi, D. Feng. "Recognition of attentive objects with a concept association network for image annotation". *Pattern Recognition*. Vol. 43. 2010. pp. 3539-3547.

5. Z. Li, Z. Shi, X. Liu, Z. Li, Z. Shi. "Fusing semantic aspects for image annotation and retrieval". *Journal of Visual Communications and Image Representation*. Vol. 21. 2010. pp. 789-805.

6. R. Kachouri, K. Djemal, H. Maaref. "Multi-model classification method in heretogeneous image database". *Pattern Recognition*. Vol. 43. 2010. pp. 4077-4088.

7. R. Fergus, F. Li, P. Perona, A. Zisserman. "Learning Object Categories Form Internet Image Search". *JPROC*. of IEEE. Vol. 98. 2010. pp. 1453-1466.

8. B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman. "Labelme: a database and webbased tool for image annotation". *IJCV*. Vol. 77. 2008. pp. 157-173.

9. A. Torralba, R. Fergus, W. T. Freeman. "80 million tiny images: a large database for nonparametric object and scene recognition". *IEEE trans. on PAMI*. Vol. 30. 2008. pp. 1958-1970.

10. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. "Learning object categories from google's image search". *Proc. of ICCV.* Vol. 2. 2005. pp. 1816-1823.

11. F. Schroff, A. Criminisi, A. Zisserman. *Harvesting Image Database from the Web*. Proc. of International Conference on Computer Vision. Rio de Janeiro, Brazil. October 2007. pp. 1-8.

12. A. Bosch, A. Zisserman, X. Munoz. *Representing shape with a spatial pyramid kernel*. Proceedings of the International Conference on Image and Video Retrieval. New York, USA. 2007. pp. 401-408.

13. A. Bosch, A. Zisserman, X. Munoz, *Image Classification using Random Forests and Ferns*. Proceedings of the International Conference on Image and Video Retrieval. Rio de Janeiro, Brazil, October 2007. pp. 1-8.

14. D. MacKay. "An Example Inference Task: Clustering". *Information Theory, Inference and Learning Algorithms*. Ed.Cambridge University Press. Chapter

20. Cambridge, United Kingdom. September 2003. pp. 284–292.

15. K. Jain. “Data Clustering: 50 years beyond K-means”. *Pattern Recognition Letters*. *Elsevier Journal*. Vol. 31. 2010. pp. 651–666.

16. N. Sasao, T. Kondo, T. Suenaga, Y. Matsumoto, T. Ogasawara. *Multi-Modal Interaction by HRP-2-Implementation of a Portrait Drawing Function*. Proceedings on Conference of Robotics Society of Japan (CD-ROM). Yokohama, Japan. September. 2005. Vol. 23. pp. 1H13.

17. S. Kudoha, K. Ogawarab, M. Ruchanurucks, K. Ikeuchi. “Painting Robot with Multi-Fingered Hands and Stereo Vision”. *Robotics and Autonomous Systems*. Vol. 57. 2009. pp. 279-288.

18. S. Calinon, J. Epiney, A. Billard. *A Humanoid Robot Drawing Human Portraits*. Proc. of IEEE-RAS Int. Conf. on Humanoid Robots. Tsukuba, Japan, December 2005. pp.161-166.