Emerging Trends in Computer Science and Information Technology (ETCSIT) 2011
Proceedings published by International Journal of Computer Applications® (IJCA)

Talking Sketch- A Sketch Generation System with LIPSYNC

Sukhwant Kaur[1], Sandhya P[2], Jayaram Kanagala, Vinayak Karande, Prathmesh Limaye

[1]Assistant Professor, [2]Lecturer
Fr.CRIT, Vashi
Navi Mumbai, Maharashtra

ABSTRACT

The computer applications available today can make an expert artist out of a layman. The existing system is a photo to sketch generator application which utilizes sketch generation algorithms found in the field of image processing. However, these applications are not friendly in terms of user interaction. Hence, despite these applications being available online on a large scale, these have not been a much of a rage among the net-savvy people and also have not caught the fancy of the industry despite offering a tremendous potential. Thus, we propose to create a system which can overcome current drawbacks by making the application more user friendly. In addition to this, we propose to implement lip-sync features to the generated sketch and make the system more engrossing.In nut shell, our system will convert digital photos to sketch and make it talk by providing lip sync features to it.

Keywords – lip sync, lip extraction, sketch generation, viseme**.**

1. INTRODUCTION

Pencil sketch drawings are a very popular form of art. In a typical pencil sketch image, only the most characteristic lines of the underlying subject are drawn, using a dark color (pencil) on a white background (paper). Also, certain degree of variation in the darkness of the pencil is typically used to depict various types of transitional boundaries (e.g., edges) and shadows in the original scene. In order to partially automate this process, tools have been developed, which allows a static, semi-abstracted cartoon image to be generated easily from photographs taken by users.

Our system focuses on converting user input image to a sketch which resembles the image and that too by restoring all the key attributes of image. Moreover, using lip sync techniques, the converted sketches can be animated to make it look more appealing to the people. Lip synchronization is the process of speech to lip motion mapping. Based on sketch conversion and lip sync techniques, a tool is created with which user can interact by loading images of his/her choices and thus provide an entertainment tool.

2. EXISTING SYSTEMS

There are few of the existing systems which are Into Cartoon Pro 3.1, PicToon, Sketch Master 4.8.

2.1 Into Cartoon is easy-to-use software for conversion of photos into cartoon or any other graphic representation. It allows the user to adjust the amount of threshold and darkness.

The drawback of this system is that the photo to be transferred to a cartoon version should have a sharp focus, or else a bad photo will bring bad effect. It can only process 24bit color BMP or JPG picture. In order to obtain higher quality sketch, it is necessary to shrink the image to a smaller size. JPG is not recommended because of color problems due to lossy compression [5]. .

2.2 PicToon is a cartoon system which can generate a personalized cartoon face from an input picture. It is easy to use and requires little user interaction.

The drawbacks of the system are that it uses the neural networks for which it requires around 200-250 training data sets which is economically infeasible in implementing the software [1]. .

2.3 Sketch Master 4.8 is a manipulation tool for the creation of realistic looking hand-drawings derived from their photos. The poor user interface to communicate with user contributes to its drawback [6].

3. PROPOSED SYSTEM

After analyzing the existing systems and their drawbacks, we propose a system named **'Talking Sketch'** which will convert digital photos to sketch and make it talk by providing lip sync features to it. In order to fulfill the above mentioned criteria, the system is divided into the following three modules.

3.1 Module 1 (Sketch generation)

This module will take a photograph of the user as an input and will make a sketch out of it.

3.2 Module 2 (Lip region extraction)

This module will extract the attributes of lip from generated sketch. The attributes of the lips are required to be extracted so as to ensure that the lip sync can be done by finding the lip region.

3.3 Module 3(Lip Synchronization)

Lip synchronization is the determination of the motion of the mouth and tongue during a speech. Intonation characteristics, a pitch, an amplitude and voiced/whispered quality, are dependent on the sound source, while the vocal tract determines the phoneme. A phoneme is the basic unit of the acoustic speech. A visual representation of the phoneme is called viseme [4]. It is the process of speech to lip motion mapping. The process of lip synchronization is divided in four main parts. In first pass, lip features are extracted using standard algorithms which use color, edge and motion information. In second pass, audio processing is done. The signal parameters are extracted. In third pass, the speech parameters are presented to classifier in two sets – Training set and Testing set.

4. DESIGN

The modular design of the project is as shown in Figure 1.

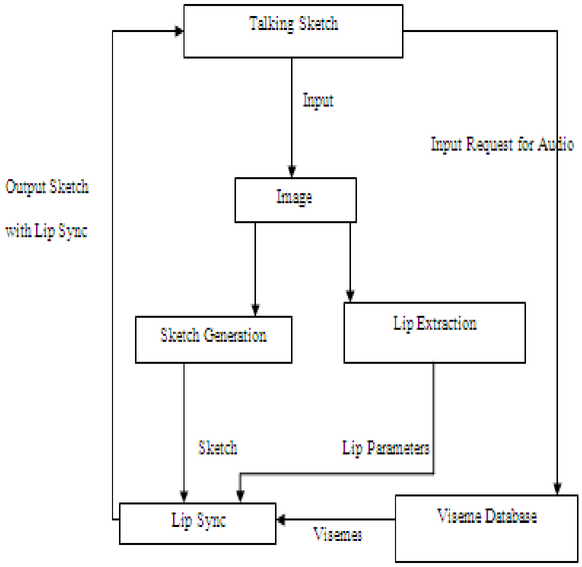

Figure 1: Module Interaction

5. PROPOSED ALGORITHMS FOR THE MODULES

5.1. Algorithm for Sketch Generation

This steps for the sketch generation algorithm are as follows:

Step 1 Smooth the input picture

In general, an unprocessed image may contain excessive noise, which reacts strongly to the subsequent gradient estimation. Therefore, as the first step, the input image is smoothed by a Gaussian low-pass filter, just as the pre-processing of the Canny's edge detector.

Step 2 Gradient Estimation

With the noise subdued by the smoothing step, the next task is to detect points of significant gradient (roughly speaking, the edges). As noted above, not only the locations of the edges but also the gradient at those locations should be kept. For illustration purpose, we use the following Laplacian operator for gradient estimation as shown in Figure 2.

0	-1	0
-1	4	-1
0	-1	0

Z_1	Z_2	Z_3
Z_4	Z_5	Z_6
Z_7	Z_8	Z_9

Figure 2: The Laplacian Mask

That is, the output after applying the mask (left) to an image block shown on the right is given as

$$g = 4z5 - (z2 + z4 + z6 + z8).$$

Since the Laplacian operator produces both positive gradient and negative gradient, to roughly detect the edge, we can simply keep either of the two gradients, a more accurate way is to detect the zero-crossing. We choose the negative gradient. To this end, the following simple thresholding is used

$$g = \begin{cases} |g| & \text{if } g < 0 \\ 0 & \text{otherwise} \end{cases}$$

Step 3 A proper Gradient Transform

To achieve the objective of linking darker pencil color to edges of larger gradient, we apply the following transform,

$$g = \begin{cases} 120 - g & \text{if } g > 0 \\ 255 & \text{otherwise} \end{cases}$$

where 120 is an empirically-chosen parameter, which can be user-adjustable in the software.

Step 4 Final smoothing for enhanced visual effects

We adopt another smoothing step to further blend the contours with the background and to link the broken contours.

Advantages

This removes the drawbacks of the previous algorithm by having a much better image quality due to the improvements made in this algorithm.

5.2. Lip extraction using distance autocorrelogram

Distance autocorrelogram is shape representation and retrieval method. Firstly, the distance autocorrelogram is obtained under the premise of getting the contour's centroidal distances. Then, we apply this shape descriptor to content-based image retrieval (CBIR). This feature depends on the centroidal distances and correlation between neighboring edges, so it can express the edge's spatial distributing information. Experimental results and algorithm analysis demonstrate the efficiency and feasibility of this shape based image retrieval approach. Distance autocorrelogram can be used to extract lip from any given image.

Step 1 Centroidal Distance Calculation
The distance autocorrelogram is obtained under the premise of getting the contour's centroidal distances.

Step 2 Shape Retrieval
Apply this shape descriptor to content-based image retrieval (CBIR). This feature depends on the centroidal distances and correlation between neighboring edges, so it can express the edge's spatial distributing information. Distance autocorrelogram can be used to extract lip from any given image [3].

Advantages

i. This method is not sensitivity to color and illumination changes.
ii. This method is invariable to translation, rotation, scaling and illumination.

5.3. The Lip Sync algorithm using Viseme databases

The steps used in this method are as follows:

Step 1 Conversion into frames
The pre recorded speech is first segmented into frames.

Step 2 Phoneme determination
For each frame most probable phoneme is determined. This is done by means of searching from the database the corresponding phoneme with respect to the sound wave of the frame.

Step 3 Viseme extraction
The phonemes extracted from the previous step are mapped onto the corresponding viseme while checking from the database.

Step 4 Viseme movement
The visemes now run simultaneously along with the sound at a rate of 25 frames per second to make it look like a continuous movement. Each of these frames are stored in the database. These frames are retrieved from the database at a fast rate so as to make it look like a moving picture. In order to have a faster access of the frames, these frames are placed in the cache [7].

Advantages
It is quite easy to implement as compared to neural networks due to the fact that large number of training data sets are not required in this algorithm.

6. TECHNICAL ISSUES CONSIDERED WHILE DESINING THE SOFTWARE

6.1 Size

The size of the image range can be 512 X 512 and higher.

6.2 Format

The formats of image can be jpg ,tif,bmp. If any other format is used then conversion needs to be done into the above mentioned formats. This is because matlab and java offers only these common formats for image processing. The java platform allows image processing using Java Advanced Imaging (JAI) API.

6.3 Lip sync issues

i. Above all it is necessary to know that the acoustic (sound) to articulatory (motion as in lip sync) mapping is not a fixed mapping. As a result of which a uniform lip sync procedure cannot be obtained. This is because each person has a way of pronunciation which might be different from others.

ii. Since we are using a lip sync ,it is necessary that we obtain the visemes from the database. For that matter there are two things arise:
   a. What type of database to be used?
   b. What will be stored in the database ?

The second point is important since the visemes may either be stored as a series of images or it may also be stored as a set of parameters of the lip location.

If the visemes are directly stored into the database then the size of the database will increase. If we select storing the parameter of the lips, then there would be complexity as in reconstructing the lip information while processing. We would prefer to go about the visemes being stored in the database even if we have a little overhead of the memory.

6.4 Problem in differentiating between eyes and lips during lip detection as both have same shape.

The image is divided into horizontal four segments.
Forehead , Eyes, Nose and Lips

The Region below the nose is therefore detected as the lip region. This can be done by using Distance Auto correlogram (DAG) algorithm. Also we can differentiate by signifying that the lips would generally be below the half portion of the face and thereby mostly find the segment in the lower half.

6.5 Issue to avoid delay in database retrieval (Viseme database) during real time play.

We have to use a buffer to store and retrieve frequent visemes from the database for fast retrieval. This buffer might be a simple use of additional variables from the java files which would store these details.

6.6 Storage of processed images

We wouldn't be storing the image because even after storing the image for playing again they need to be processed. Moreover we assume that the retrieval of the stored processed image is more or less going to take the same time as that of processing the input image. So there isn't much difference in the overhead as such.

6.7 Selection of language for implementation

Flash action script 2.0 and JAVA are 2 languages available. Though the action script language is simple with inbuilt functions it doesn't allow connectivity with Matlab, which does the image processing. Whereas JAVA is more user interactive and has very efficient compatibility with MATLAB.

6.8 Memory requirements

It depends on how much visemes need to be generated for particular session. There are different standards for visemes selections. Out of which we would be selecting Mpeg 4 standards which has 14 visemes. In the English language there are 40 phonemes which are being mapped onto 14 visemes and it is one to many mapping depending on the requirement of sentence spoken. That is each phoneme may have many visemes to be mapped onto.

7. CONCLUSION

We have tried to put forth the **'Talking Sketch'**, a sketch generation system that creates a personalized cartoon from an input image. We have researched on various tools and techniques to see to it that the sketch generated by using our system should look somewhat artistic and must resemble the original person. The system should also be easy to use among the people. A lip sync algorithm has also been researched that will make the generated sketch 'talk' more easily and naturally. We look forward to implement the algorithms researched and listed in this paper to make our project successful.

8. REFERENCES

[1] Hong Chen, Nan-Ning Zheng, Lin Liang, Yan Li, Ying-Qing Xu, Heung-Yeung Shum, 'PicToon :A Personalized Image-based Cartoon System' Xi'an Jiaotong University, Microsoft Research, Asia

[2] Jin Zhou and Baoxin Li., 'Automatic Generation of Pencil-Sketch like drawings from Personal Photos', Center for Cognitive Ubiquitous Computing, Department of Computer Science and Engineering, Arizona State University, Tempe, AZ, U.S.A

[3] Jie-xian Zeng, Yong-gang Zhao, Xiang Fu, 'A Novel Shape Representation and Retrieval Algorithm: Distance Autocorrelogram', Journal of software, vol. 5, no. 9, September 2010

[4] Mahesh Goyani, Namrata Dave, Narendra Patel, 'Performance Enhancement in Lip Synchronization Using MFCC Parameters' International Conference on Computational Intelligence and Communication Networks (CICN-2010), 26-28 Nov. 2010, Bhopal