RECOGNITION AND TRACKING MOVING OBJECTS USING MOVING CAMERA IN COMPLEX SCENES

Archana Nagendran[1], Naveena Dheivasenathipathy[2],Ritika V. Nair[3] and Varsha Sharma[4]

Department of Information Technology, Amrita School of Engineering, Coimbatore, India.

ABSTRACT

In this paper, we propose a method for effectively tracking moving objects in videos captured using a moving camera in complex scenes. The video sequences may contain highly dynamic backgrounds and illumination changes. Four main steps are involved in the proposed method. First, the video is stabilized using affine transformation. Second, intelligent selection of frames is performed in order to extract only those frames that have a considerable change in content. This step reduces complexity and computational time. Third, the moving object is tracked using Kalman filter and Gaussian mixture model. Finally object recognition using Bag of features is performed in order to recognize the moving objects.

KEYWORDS

Key frame, stabilization, motion tracking, recognition, bag of features.

1. INTRODUCTION

Computer vision has gained paramount significance in recent times due to the increased use of cameras as portable devices and their incorporation in standard PC hardware, mobile devices, machines etc. Computer vision techniques [1] such as detection, tracking, segmentation, recognition and so on, aim to mimic the human vision system. Humans hardly realize the complexities involved in vision, but in fact, our eye is more powerful than it seems. It processes around 60 images per second, with each image consisting of millions of points. Computer vision is still a long away from its goal of replicating the human eye, but in the meantime various computer vision techniques are being applied to complex applications.

Determining whether an image contains some specific object, feature, or activity is a common problem that is dealt with. The existing methods [4] for dealing with this problem are capable of solving it only for specific objects, such as human faces, vehicles, characters, printed text etc. and in specific situations, with well-defined pose of the object relative to the camera, background, and illumination.

In this paper, we deal with the recognition of a variety of moving objects which may be present in dynamic backgrounds. The proposed algorithm is resistant to small illumination changes and also involves a module that reduces effects of camera movement.

DOI : 10.5121/ijcsea.2014.4203

2. VIDEO STABILIZATION

A moving camera may either be attached to a vehicle or may be handheld. Videos captured from handheld cameras are very jittery. Those attached to vehicles may also undergo jitter and sudden changes. It is highly undesirable to work with such video frames since it is tedious to differentiate between the foreground and the background. Hence stabilizing the video before further processing becomes a necessity.

Usually videos are stabilized by tracking a prominent feature that is common to all the frames and using it as an anchor point to cancel out all disturbances relative to it. But to implement such a method, we are required to know the position of the prominent feature in the first frame. In this paper, we explain a method of video stabilization which does not require such presumptive knowledge, but rather uses a method of point feature matching which is capable of automatically searching for the background plane in a video sequence and using its observed distortion to correct for camera motion.

The basic idea of the proposed stabilization algorithm is to first determine the affine image transformations between all neighbouring frames of the video by using a Random Sampling and Consensus (RANSAC) procedure [3] applied to point correspondences between two images. Then the video frames are warped to achieve a stabilized video.

The algorithm consists of the following steps. Initially we read the first two video frames, say frame A and frame B. They are read as intensity images (since colour is not necessary and also because using grayscale images improves speed) and points of interest from both frames are collected, preferably the corner points of all objects in the frame. Then we extract features for each set of points and find likely correspondences between both the set of points. The matching cost used between the points is the sum of the squared differences (SSD) between their respective image regions. Since we do not apply any uniqueness constraint, points from frame B can correspond to multiple points in frame A. To get rid of incorrect correspondences and to obtain only the valid inliers, we make use of the RANSAC (Random Sample Consensus) algorithm. Next we find the affine transform between the points of frame A and frame B. It is a 3-by-3 matrix of the form:

$$\begin{bmatrix} a_1 & a_3 t_r \\ a_2 & a_4 t_c \\ 0 & 0 \quad 1 \end{bmatrix}$$

This transform can be used to warp all the succeeding frames such that their corresponding features will be moved to the same image location. The cumulative distortion of a frame relative to the first frame will be the product of all the preceding inter-frame transforms.
For numerical simplicity, we re-fit the above affine transform into a simpler scale-rotation-translation transform, which has only four free parameters: one scale factor, one angle, and two translations. This s-r-t transform matrix is of the form:

$$\begin{bmatrix} s*\cos(\theta) & s*-\sin(\theta) & t_x \\ s*\sin(\theta) & s*\cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

The above steps are iteratively applied to all the video frames, hence resulting in a smooth and stabilized video sequence.

3. KEY FRAME EXTRACTION

Key frame is the frame which indicates the content of the video. We use a key frame selection technique in our algorithm so as to avoid unnecessary processing on unimportant frames hence saving us time and memory space. Many methods that are discussed for key frame extraction are histogram method, template matching, pixel-based comparison [6, 7]. In pixel-based comparison, each pixel is compared,so as tokeep the time complexity high.In the histogram method, the location information is entirely lost. Two images will have different content but similar histograms. Hence, we use edge based method to consider the content of the frames.

The proposed approach for key frame selection method is to compute the edge difference between two consecutive frames. The frame which exceeds the threshold is considered as the key frame. The key frames are selected to find the important frames for describing the content of the video for later processes. The edge difference is computed, because it is edge dependent.

The proposed method is elaborated below. For i=1 to N, where N is the total number of frames in the video,

i. Read frame V_i and V_{i+1}. Find grayscale image for V_i and V_{i+1}. Let G_i be the grayscale image of V_i and G_{i+1}be the grayscale image of V_{i+1}. Find the edge difference between these two grayscale images G_i and G_{i+1}.
ii. Compute the mean and standard deviation.
iii. Find the threshold value. The threshold is computed as:
Threshold = $M + a * S$
where$M$ is the mean, a is a constant and S is the standard deviation.
iv. Compute the key frames for the video from i=1 to (N-1) as:
If diff(i) is greater than the given threshold value then write V_{i+1}as the output frame, otherwise check for rest of the frames in the video.

4. OBJECT DETECTION AND TRACKING

The next step in our proposed procedure involves object detection and tracking. The Gaussian mixture model [8] and Kalmanfilter [9] have been used to perform the same.In this paper, motion based object tracking is divided into two parts:first, detecting moving objects in each frame. Second, associating the detections corresponding to the same object over time. A background subtraction algorithm based on Gaussian mixture models is used forthe detection of moving objects and noise elimination is done by applying morphological operations to the resulting foreground mask. In the Gaussian Mixture model, frame pixels are deleted from the required video to achieve the desired results. Thebackground modeling, using the new video frames, calculates and updates the background model. The main reason behind the use of a background model is that it should besensitive enough to identify all moving objects of interest as well as robust against environmental changes in the background.

Foreground detection compares the video frame with the background model, and identifies foreground pixels from the frame. The approach used here for foreground detection isto check whether the pixel is significantly different from the corresponding background estimate. To improve the foreground mask based on the information obtained from the outside background model, data validation is performed.In order to help the tracker detect motion, we extract the background image from sequences of frames. To extract the background image from a sequence of frames, every pixel of the background image is separately calculated using the mean or the median or the highest appearance frequency value from the series of frames. This results in a difference image.In order to avoid noise, the difference image iscompared with the threshold

value.The threshold value is determined based on the highest pixel differences after sampling of many frames.

Finally, blob analysis is performed, which consequently detects groups of connected pixels, which correspond to the moving objects. The attribute considered here for tracking is 'motion'. All motion pixels in the difference image are clustered into blobs. We have used an updated version of the popular image processing "fill" procedure; the implementation can extract blobs either from the whole image or from a specific part of it.AKalman filter is designed for tracking, which is used to predict the object's future location,to reduce noise in the detected location and to help associate multiple physical objects with their corresponding tracks.

Based on the fact thathigh level semantic correspondences are indispensable to make tracking more reliable, a unified approach of low-level object tracking and high-level recognition is proposed here for single object tracking, where the target category is recognized during tracking.
Track maintenance is an important aspect .In any given frame, somedetections and tracks may remain unassigned detections while other detections may be assigned to tracks. Using the corresponding detections the assigned tracks are updated andthe unassigned tracks are marked invisible. Theunassigned detection is taken as a new track. Each track keeps count of the number of consecutive frames. If the count exceeds a specified threshold value, we assume that the object has left the view and it deletes the track. The object of interest is initialized by a user-specified bounding box and the tracks are numbered.

5. OBJECT RECOGNITION

Object recognition in computer vision is the task of finding a given object in an image or video sequence. The method used in this paper, object recognition using Bag of features [10, 11] is one of the successful methods for object classification. The basic principle of object recognition using Bag of features states that every object can be represented using its parts. Thus, the parts of the objects are recognized and then the objects are classified based on these parts. There are four main steps in this method:

i. Feature extraction
ii. Learning visual vocabulary
iii. Feature quantization using visual vocabulary
iv. Image representation

Initially, the corners in the image are found using the Harris corner detection [12] technique. Now we calculate the Scale Invariant Feature Transform (SIFT) features [13, 14] (128 dimension vector) around each corner point. These vectors represent the parts of the object that have to be recognized. Next, we form a dictionary by taking different objects and also different images of the same object from different angles and then train the dictionary. For this, we repeat the first step i.e. corner detection and feature extraction for each of the images and store the 128 dimension vector (for SIFT) for every corner detector in the image into an array. Thus, a large matrix will be formed. Using this matrix, clustering is done among the data using the K-means [15] clustering method. Each of the cluster centres are taken as a representation for each part. Thus, we now have a dictionary which contains different parts which are represented as cluster centres. In order to find the frequency of parts in an object, the initial step i.e. corner detection and SIFT feature calculation, is performed and the nearest parts matching these SIFT features are found. So every SIFT feature is categorized into one of the parts based on the distance from the SIFT feature to the cluster centres. With the cluster centres as x-axis and frequency on y-axis we form a histogram[16] that represents the frequency of parts (Figure 1). Thus, every image is represented with a histogram which depicts the frequency of the parts. These histograms are

matched and hence object recognition is accomplished.

The above explained technique of object recognition requires two set of images: First is the training set, to which objects are matched. Second is the test set, which contains objects that need to be recognized. The images in the training set are trained and the histograms are stored beforehand, while the images in the test set need to be trained and the histogram of parts have to be calculated and then matched with the trained set of images and then recognized.

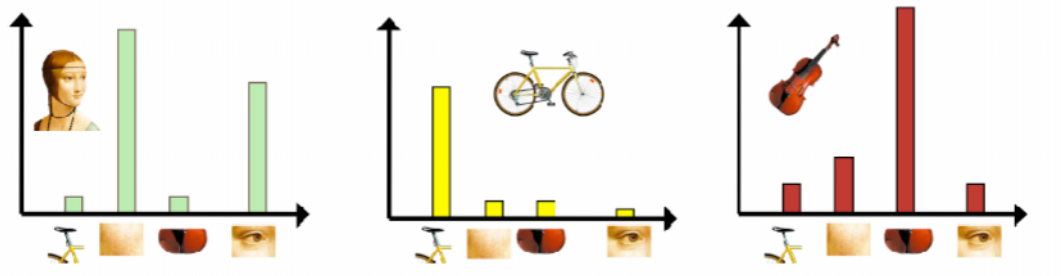

Figure 1. Histogram representation of an object

6. CONCLUSION

This paper "Recognition and tracking of moving objects using moving camera in complex scenes" is to detect and recognize the moving objects in complex backgrounds using various different techniques. Various videos having complex backgrounds have been evaluated and the above methods successfully detect and recognize mainly four classes of objects and produces good recognition results. With the increment in the number of images in the database the computational time also increases by a small value.

Future work includes reduction in computational time as well as increasing recognition for more number of categories.

ACKNOWLEDGEMENTS

The authors ArchanaNagendran, NaveenaDheivasenathipathy,Ritika V. Nair and Varsha Sharma would like to thank Ms. G. Radhika, Ms.Aarthi R. and Ms.Padmavathi S. for their guidance and useful comments.

REFERENCES

[1] GerardMedioni and Sing Bing Kang, Emerging topics in computer vision, Prentice Hall, 2004.
[2] GottipatiSrinivasBabu, "Moving object detection using Matlab", IJERT, vol. 1, issue 6, August 2012.
[3] Marco Zuliani, RANSAC for Dummies, August 2012.
[4] Byeong-Ho Kang, "A review on image and video processing", vol. 2, International Journal of Multimedia and Ubiquitous Engineering, April 2007.
[5] KhushbooKhurana and Dr. M. B. Chandak, "Key frame extraction methodology for video annotation, IJCET, vol. 4, issue 2, March-April 2013, pp. 221-228.
[6] C.F.Lam, M.c.Lee,"Video segmentation using colour difference histogram", Lecture Notes in Computer Science, New York: Springer Press, pp.159-174, 1998.
[7] D. Borth, A. Ulges, C. Schulze, T. M. Breuel, "Key frame Extraction for Video Tagging & Summarization", volume S-6 of LNI, page 45-48, 2008.

[8] D. Hari Hara Santosh, P. Venkatesh, P. Poornesh, L. NarayanaRao, N.Arun Kumar,"Tracking Multiple Moving Objects Using Gaussian Mixture Model",International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-2, May 2013.
[9] Hitesh A Patel1, Darshak G Thakore2, "Moving Object Tracking Using Kalman Filter",IJCSMC, Vol. 2, Issue. 4, April 2013, pg.326 – 332.
[10] HHerv´eJ´egou, MatthijsDouze, and CordeliaSchmid. Improving bag-of-features for large scale image search. International Journal of Computer Vision, 87(3):316–336, 2010.
[11] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In CIVR, pages 494–501, 2007.
[12] Chris Harris and Mike Stephens, A Combined Corner and Edge Detector, Proceedings of The Fourth AlveyVision Conference (Manchester, UK), pp. 147-151, 1988.
[13] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal Of Computer Vision, 60(2):91–110, 2004.
[14] David G. Lowe. Object recognition from local scale-invariant features. In ICCV, pages 1150–1157, 1999.
[15] KhaledAlsabti , Sanjay Ranka, Vineet Singh, "An Efficient K-Means Clustering Algorithm".
[16] E. Hadjidemertriou, M. Grossberg,and S. Nayar. Multiresolution histograms and their use in recognition. IEEE Trans. PAMI, 26(7):831-847, 2004

Authors


All four authors are students of Amrita School of Engineering, Coimbatore, India.

We are currently pursuing our B.Tech degree in Computer Science and Engineering.


Our areas of interest are image processing and computer vision. We are currently working on an obstacle recognition aid for the visually impaired using the techniques elaborated in this paper.