التعداد الدقيق للحجاج أثناء النفرة من عرفات إلى المزدلفة ورمي الجمرات بمنى

د. رغيد محمد عطا

كلية الهندسة، جامعة طيبة، المدينة المنورة

د. ياسين محمود ياسين

كلية المجتمع، جامعة طيبة، المدينة المنورة

الهدف الأساسي من تقنيات معالجة الصور هي استخراج معلومات هامة من بيانات الصورة. وحيث أن استخراج مثل هذه المعلومات بالعين المجردة يستغرق وقت وجهد وتكلفة عالية لذلك نلجأ إلى استخدام الكمبيوتر للقيام بمثل هذا العمل. فمثلا إحصاء الناس في منطقة مزدحمة دون استخدام مثل هذه التقنيات المتقدمة هي مهمة شاقة تستغرق وقتا طويلا وتعطي نتائج لا يمكن الاعتماد عليها كلية بسبب العديد من العوامل مثل ظروف الإضاءة الصعبة لمكان تجمع الناس، وحركة الناس الغير منتظمة والفقدان التدريجي لتركيز الشخص الذي يقوم بمهمة العد والإحصاء. وبالتالي، فإن تطوير هذا الهدف له أهمية قصوى في تطبيقات كثيرة متنوعة وخاصة في مناسبات مثل الحج.

في هذا البحث تم اقتراح وتطوير طريقة تلقائية لعد واحصاء الحجيج اعتمادا على تقنية معالجة الصور والعد الآلى المعتمد على الرؤية التلقائية مع خطط فعالة للمعالجة المسبقة الضرورية للصورعن طريق توصيل مجموعة من الكاميرات الرقمية لاسلكيا بالحاسب الآلى لتمكننا من التقاط مجموعة من الصور لنفس الهدف من زوايا مختلفة. زمن ثم يتم تجزئة واستخراج وتحليل وتصنيف الملامح مع بناء قاعدة بيانات من التصنيفات واستكشاف مختلف التدابير والتنوع فى اتخاذ قرار التصنيف بقدر الحاجة. ولكي يتم بناء قاعدة بيانات تساعد فى التصنيف أثناء العد تم النظر في مجموعة متنوعة من التكتلات البشرية في ظروف إضاءة مختلفة وخلفيات متعددة من حالات التزاحم، وحالات متنوعة من كثافات مختلفة من حشود من الناس، وأنواع وسرعات مختلفة من تحركات الناس، والسمات المميزة التي يمكن انتقائها على أساس الشكل أو الحجم أو لون الجلد. ومن ثم تم تحويل مجموعة البيانات التي تم جمعها من الصور الملتقطة إلى ملامح يمكن تميزها. وبذلك تكون النتيجة النهائية لهذه الخطوة هو توصيف مجموعة من السمات المعروفة باسم مجموعة الملامح، التي تعبر بدورها عن مجموعة البيانات من الصور الملتقطة. وهذه الخطوة الوصفية يمكن أن تنطوي على مجموعة من الأنشطة المختلفة، ولكنها مترابطة، مثل: **اختيار الملامح النهائية ، المعالجة الأولية، الجدولة، استخراج الملامح ثم التصنيف.** وبالتالي، فإن النظام الآلي لعد الحجيج المقترح تنفيذه يستخدم منهج توصيف قوي، ومجموعة من التصنيفات داخل حزمة التوصيف الموحد.

وبالتالي فهذه الطريقة يمكن الإعتماد عليها فى عد واحصاء الحجيج أثناء انتقالهم فى المشاعر سواء للذهاب إلى الجمرات أو الإنتقال من عرفة إلى مزدلفة بدقة ودون مشقة، مما يساعد في تحسين عمليات الفحص والتتبع، والسماح للكشف المبكر لحالات الإزدحام التي تهدد الحياة بشكل كبير مما يساعد في اتخاذ الإجراءات المناسبة في الوقت المناسب.

Accurate counting Method for Pilgrims During their flow from Arafat To Muzdalifa and Mina

Raghied Mohammed Atta[1] and Yaseen Mahmoud Yaseen[2]

[1]Faculty of Engineering, [2]Community College, Taibah University, Madinah, KSA

Abstract

In this paper we present a real-time crowd model used for counting pilgrimage crowds during their movements from Arafat to Mozdalifa and further to Mina based on continuum dynamics. In this model, a dynamic potential field simultaneously integrates global navigation with moving obstacles such as other people, efficiently solving for the motion of large crowds without the need for explicit collision avoidance. Simulations created with this system run at interactive rates, demonstrate smooth flow under a variety of conditions, and naturally exhibit emergent phenomena that have been observed in real crowds. We focus on real-time synthesis of crowd motion based on continuum dynamics for thousands of individuals with intersecting paths. Our formulation is designed for large groups with common goals, not for scenarios where each person's intention is distinctly different.

1. Introduction

The estimation of the number of people present in an area can be an extremely useful information both for security/safety reasons especially for Haj period (for instance, an anomalous change in number of persons could be the cause or the effect of a dangerous event) and for economic purposes (for instance, optimizing the schedule of public transportation system on the basis of the number of passengers). Hence, several works in the fields of video analysis and intelligent video surveillance have addressed this task.

Human crowds are ubiquitous in the real world, making their simulation a necessity for realistic interactive environments. The crowd counting problem can be classified into two tasks: crowd counting across a detection line in certain time duration (line of interest (LOI) counting), and estimating the total number of pedestrians in some region at each time (region of interest (IRO) counting).
In the published literature, there are two classes of methods for ROI counting; feature and pixel regression, and pedestrian detection. Feature or pixel regression methods extract the feature vectors in the region of interests and use the machine learning algorithm to regress the number of pedestrians from number of features and pixels in foreground blobs or segmented motion segments. The features include edges [1], wavelet coefficients [2], or combination of a large bank of features [3,4]. The regression method may be linear regression [1], neural network [5], Gaussian process regression [4] or discrete classifier [3]. The number of features carried by one pedestrian is heavily affected by detecting camera perspective. The group count estimation greatly depends on the quality of the background subtraction. The better the foreground masks returned, the more accurate the count estimates are. So they always need retraining using large

amount of annotated data from the specific scene, which makes it inconvenient to deploy in practical applications. Physically correct crowd models also have applications outside of computer graphics in psychology, transportation research, and architecture.

Pedestrian detection methods count pedestrians by multi-target detection. The detection is completed by background differencing [6], motion and appearance joint segmentation [7], silhouette or shape matching [8], or standard object recognition method. Though there are great progresses in object detection in recent years, robust detection of pedestrian under crowd environment is still a challenging problem. The performance of pedestrian detection method will decrease rapidly when the crowd density and occlusion degree increase, as in our case. Feature regression and pedestrian detection are only applied to ROI detection. For LOI counting, most literature adopt feature tracking. Features across frames are tracked into trajectories, and the trajectories are clustered into object tracks. Examples include [9-14]. The tracking based methods are hardly robust under crowd environment, and their time consumption is often huge for real-time systems.

Real-time crowd simulation is difficult because large groups of people exhibit behavior of enormous complexity and subtlety. A crowd model must not only include individual human motion and environmental constraints such as boundaries, but also address a bewildering array of dynamic interactions between people. Further, the model must reflect intelligent path planning through this changing environment. Humans constantly adjust their paths to reflect congestion and other dynamic factors. Even dense crowds are characterized by surprisingly few collisions or sudden changes in individual motion. It has proven difficult to capture these effects in simulation, especially for large crowds in real-time.

Most work has been agent-based, meaning that motion is computed separately for each individual. The agent-based approach is attractive for several reasons. For one, real crowds clearly operate with each individual making independent decisions. Such models can capture each person's unique situation: visibility, proximity of other pedestrians, and other local factors. In addition, different simulation parameters may be defined for each crowd member, yielding complex heterogeneous motion. However, the agent-based approach also has drawbacks. It is difficult to develop behavioral rules that consistently produce realistic motion. Global path planning for each agent quickly becomes computationally expensive, particularly in real-time contexts. As a result, most agent models separate local collision avoidance from global path planning, and conflicts inevitably arise between these two competing goals. Moreover, local path planning often results in myopic, less realistic crowd behavior. These problems tend to be exacerbated in areas of high congestion or rapidly changing environments.
This paper presents a real-time motion synthesis model for large crowds without agent-based dynamics. We view motion as a perparticle energy minimization, and adopt a continuum perspective on the system. This formulation yields a set of dynamic potential and velocity fields over the domain that guide all individual motion simultaneously. Our approach unifies global path planning and local collision avoidance into a single optimization framework. People in our model do not experience a discrete regime change in the presence of other people. Instead, they perform global planning to avoid both obstacles and other people. Our dynamic potential field

formulation also guarantees that paths are optimal for the current environment state, so people never get stuck in local minima.

2. Techniques and Algorithms

Our goal is to calculate bounds for the count and location of people in an area from a planar projection of the visual hull. The first step is to compute this projection from the silhouettes measured by the sensors through background subtraction. The projection is a set of polygons. The second step is to compute bounds to the number of objects in each polygon. As objects move, these bounds change and can be improved over time. A tree is used to record their history. Finally, the tree and its associated polygons are used to localize those workspace regions that are occupied by people.

People walking from Arafat towards Mina move along a plane. Therefore, it is only necessary to project the visual hull onto this plane. The projection contains the information from the 3D visual hull that is most useful for counting and localizing people. When people walk reasonably close to each other, or occlude each other respective to the camera frame, the foreground blobs corresponding to these people merge together to produce a single blob. This overlap of different foreground targets makes an individual count extremely challenging. The effect is even more detrimental when there are large groups of people walking together, which is our case. This is the primary motivation for our work.

2.1 Image Pre-Processing

Before the image undergoes the process of segmentation, the image acquired from the camera has to be pre-processed. The resolution of the image processed by the system must be fixed regardless of the resolution of the image. Two criteria must be considered; first, If the resolution of the image is too large, noise become dominated and the image is blur and the processing time will be slower as more pixels has to be handled by the computer. Second, if the resolution of the image is too small, it is so difficult to see and represent the information. The image handled by the system is monochrome image. As a result, 8-bit bitmap format is used instead of 24-bit RGB image. The processing time will be faster as the computer handles fewer bits.

2.2 Cluster Segmentation

Segmentation is an image processing technique which is used to extract the object from the background in an image. Each object occupies certain pixels in the image. Generally, there are some differences in the pixel value and gray level of the pixels belonging to the object with reference to the pixels of the background image.

In the system, the absolute difference of pixel values between the input image and a reference background image is used for segmentation. Also, the detection of the difference of pixel value between the input image and the reference background image is based on a sub-block of pixels rather than individual pixels. This is due to the texture of the floor of the counting zone containing some noise pixels in the output image.

In the segmentation process, these noise pixels can be eliminated when a sub-block of pixels is treated as a unit. Since there are n^2 pixels in each sub-block, it is unlikely that all pixels in a sub-block are noise pixels. As a result, the noise effect of the total pixels intensity of a sub-block is relatively much smaller than that of an individual pixel. If the sub-block of pixels is found that it is not occupied by the object, then the whole sub-block will be changed to white pixels including the noise pixels in the sub-block. Also, the total pixel value of the sub-block will be compared with the sub-block at the same position of the reference background image. If a sub-block of the input image has noise pixels, the sub-block of the reference background image at the same position will also has noise pixels. The noise pixels will then compensate each other during comparing and subtraction of the pixel value of the two sub-blocks.

2.3 Sub-Block Processing

Initially, the counting zone of the image is divided into several sub-blocks. In the system, each sub-block consists of nxn pixels. Let us focus on one sub-block in the counting zone. The n^2 pixel values of the pixels of the sub-block will be added together and then the total pixel value of the sub-block can be obtained. The total pixel value of the sub-block in the image will then be compared with the sub-block of the background image at the same position.

Normally the gray-level of the sub-block of the processing image is slightly darker than that of the background image due to the shadow of the people but it can be seen that if the sub-block of the input image belongs to the background, the difference in the gray level and the total pixel value of the two sub-blocks will not be very large due to similar pixel value and gray level. As a result, if the sub-block of the input image belongs to the background, then the absolute difference of the total pixel value between the sub-block of the input image and the sub-block of the background image will not be very large.

As before, the pixel value of the n^2 pixels of the sub-block will be added together and the total pixel value of the sub-block is obtained and then the total pixel value will be compared with the ones of the sub-block of the background at the same position**.** The absolute difference of the total pixel value between the two sub-blocks will be very large. As a result, if the absolute difference of the total pixel value of the two sub-blocks is very large and greater than a threshold value, the sub-block will be considered to belong to the object and all the pixels of the sub-block will be changed to black pixels.

If the absolute difference of the total pixel value of the two sub-blocks is small and less than a threshold value, the sub-block will be considered as belonging to the background and all the pixels of the sub-block will be changed to white pixels. Whether the sub-block of the input image is changed to a block of white pixels to represent the background or changed to a block of black pixels to represent the object depends on the threshold value. If the absolute difference of the total pixel value between the sub-block of the input image and the sub-block of the background image is large and greater than the threshold value, the sub-block of the input image will be changed to a block of black pixels to represent the object. Otherwise the sub-block of the input image will be changed to a block of white pixels to represent the background.

3. Discussions and Results

The first problem addressed is the effect of perspective, which causes that the farther the person is from the camera, the fewer are the detected interest points. In order to account for this effect, we need to computes the distance of each person or group of persons from the camera. To obtain this information, we first partition the detected points into groups corresponding to different groups of people. This can be treated as a clustering problem, but with the peculiarity that the shape of the clusters, their number and their densities are not known a priori. Because of this, commonly used clustering algorithms such as *k-means* and *DBSCAN* cannot be applied. However, other graph-based clustering algorithms such as Foggia [15] can perform this task, which provides a good partitioning when the clusters are reasonably separated, without requiring any a priori information about the clusters.

Once the detected points are divided into clusters, the distance of each cluster from the camera is derived from the position of the bottom points of the cluster applying an Inverse Perspective Mapping (IPM). The IPM is based on the assumption that the bottom points of the cluster lie on the ground plane.

Another factor that has to be taken into account is the effect of people density in a group. The more the persons in a group are close to each other, the more partial occlusions occur, reducing the visible part of the body, and thus the number of interest points per person. To consider this effect we need to compute a rough estimate of the people density by measuring how close the interest points in the group are. More precisely, we need to measure the ratio between the number of interest points in the group and the area covered by the group itself.

Given the need to consider not only the number of points, but also the distance from the camera and the density, the relation between these measurements and the number of people cannot be a simple direct proportionality as in Albiol's method [16]. Actually, even if a single measurement were involved, the relation might have been non linear, at least in principle; with three measurements, there is the problem of understanding their relative weights and how they interact with each other to determine the count estimate.

Since this problem cannot be easily solved analytically, we can choose to learn this relation by using a trainable function estimator. More precisely, by using a variation of the Support Vector Machine known as ϵ-*Support VectorRegressor* (ϵ-SVR for short) as function estimator. The ϵ-SVR receives as its inputs the number of points of a cluster, the distance from the camera and the point density of the cluster, and is trained (using a set of training frames) to output the estimated number of people in the cluster. The ϵ-SVR is able to learn a non linear relation and shows good generalization ability.

A further problem that is addressed here is the stability of the detected interest points. The points found by the Harris corner detector are somewhat dependent on the perceived scale and orientation of the considered object: the same object will have different detected corners if its image is acquired from a different distance or when it has a different pose.

To mitigate this problem we can adopt the SURF algorithm proposed in [17]. SURF is inspired by the SIFT scale-invariant descriptor [18], but replaces the Gaussian-based filters of SIFT with filters that use the Haar wavelets, which are significantly faster to compute. The interest points found by SURF are much more independent of scale (and hence of distance from camera) than the ones provided by Harris detector. They are also independent of rotation, which is important for the stability of the points located on the arms and on the legs of the people in the scene.

As with the Albiol's method, the output count is passed through a low-pass filter to smooth out oscillations due to image noise.
Thus, an outline of the proposed method is composed by the following steps:
1 the SURF interest points of the current frame are computed;
2 the motion vectors of the interest points are calculated by block matching between current and previous frame; the points whose speed is under a threshold are removed;
3 the remaining points are partitioned into clusters; for each cluster the distance from the camera and the density are estimated;
4 the number of points, distance and density of each cluster as given as an input vector to the ϵ-SVR regressor; the sum of the regressor outputs over all the clusters gives the initial estimate of the number of people;
5 the initial estimate is averaged over a moving window of multiple frames in order to obtain the system output.

4. Conclusions and Future Work

In this paper we propose a dynamic potential field simultaneously integrates global navigation with moving obstacles such as other people, efficiently solving for the motion of large crowds without the need for explicit collision avoidance. The system should run at interactive rates with smooth flow under a variety of conditions, and naturally exhibit emergent phenomena that have been observed in real crowds.

The group count estimation greatly depends on the quality of the background subtraction. The better the foreground masks returned, the more accurate the count estimates are. If the foreground blobs represent the foreground objects of interest well, the system returns a much better count estimate of the people in the image. When we are unable to give a good background subtraction (due to the nature of the scene) as input to the count estimation, it is difficult to obtain a good count estimate.

Most of our future work will revolve on trying the system during Haj period and enhance the background subtraction. Shadow removal improves quality of the foreground masks returned. The continuous re–learning of the background by the layering algorithm makes it adaptive to the scene conditions such as illumination changes. However it still needs an empty frame during initialization for identifying the background, but this is the case for almost any background subtraction algorithm.

We will try the Infrared Cameras instead of normal video camera to enhance the edges and hence the image recognition.

References

[1] A. C. Davies, J. H. Yin, and S. A. Velastin. Crowd monitoring using image processing. Electronics and Communication Engineering Journal, 1995.
[2] S. Lin, J. Chen, and H. Chao. Estimation of number of people in crowded scenes using perspective transformation. IEEE Trans. on Systems, Man, and Cybernetics, 2001.
[3] A. Marana, L. da Costa, R. Lotufo, and S. Velastin. On the efficacy of texture analysis for crowd monitoring. In Proc. Computer Graphics, Image Processing and Vision, 1998.
[4] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In CVPR, 2008.
[5] S. Y. Cho, T. W. S. Chow, and C. T. Leung. A neural-based crowd estimation by hybrid global learning algorithm. IEEE Tran. on Systems, Man, and Cybernetics, 29(4), 1999.
[6] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami. Fast crowd segmentation using shape indexing. In ICCV, 2007.
[7] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. IJCV, 63(2), 2005.
[8] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination
[9] V. Rabaud and S. Belongie. Counting crowded moving objects.
In CVPR, 2006.
[10] G. Antonini and J. P. Thiran. Counting pedestrians in video sequences using trajectory clustering. IEEE Trans. on Circuits and Systems for Video Technology, 16(8), 2006.
[11] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In CVPR, 2006.
[12] B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In ICCV, 2007.
[13] A. Albiol, I. Mora, and V. Naranjo. Real time high density people counter using morphological tools. IEEE Trans. On Intelligent Transportation Systems, 2(4), 2001.
[14] T. Chen, T. Chen, and Z. Chen. An Intelligent People-Flow Counting Method for Passing Through a Gate. In Robotics, Automation and Mechatronics, IEEE Conf. on, 2006.
[15] P. Foggia, G. Percannella, C. Sansone, and M. Vento. A graphbased algorithm for cluster detection. International Journal of Pattern Recognition and Artificial Intelligence, 22(5), 2008.
[16] A. Albiol, M. Silla, A. Albiol, and J. Mossi. Video analysis using corner motion statistics. In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009.
[17] H. Bay, A. Ess, T. Tuytelaars, and L. Gool. Surf: Speeded up robust features. Computer Vision and Image Understanding, 110(3), 2008.
[18] D. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 2004.