Weina Ge and Robert T. Collins

Crowd Density Analysis with Marked Point Processes

The abundance of video footage from surveillance systems in public spaces has become a driving force for advances in crowd image analysis. Of particular interest is crowd density analysis, where the goal is to detect and count people in a crowded scene. This is a challenging problem for a human observer when large numbers of constantly moving individuals are present. It is therefore desirable to have computational assets that can assist security personnel for real-time crowd monitoring. Automated analysis holds the potential to increase situational awareness for crowd control and public safety by providing real-time estimates of the number of people entering or exiting a venue.

This article presents a Bayesian approach that estimates the count and location of individuals in a video frame. Crowds are modeled by a marked point process (MPP) that couples a spatial stochastic process governing number and placement of individuals with a conditional mark process for selecting body size, shape, and orientation. Given a noisy, binary mask image where pixels are labeled foreground or background, the approach seeks a configuration of cutout shapes that simultaneously "covers" as many foreground pixels and as few background pixels as possible. We use reversible jump Markov chain Monte Carlo (RJMCMC) to search a combinatorial space of varying numbers and locations of people and to estimate the most probable configuration.

Digital Object Identifier 10.1109/MSP.2010.937495

CROWDS AS A MARKED POINT PROCESS

Unlike Markov random field models that predefine a fixed number of nodes and links between the nodes, point processes offer a more flexible framework for dealing with dynamic scenes where varying numbers of people are constantly moving in and out of view. An MPP [1] is a stochastic process that models a random number of objects randomly distributed in a bounded region with attributes (such as shape appearance) controlled by random parameters. Our notion of shape appearance is decomposed into two parameter sets: an extrinsic shape mapping and a set of intrinsic shape classes [2]. The extrinsic shape mapping determines the translation, rotation, and scaling of a centered shape model into image pixel coordinates. The intrinsic shape classes specify a library of different reference shape prototypes (e.g., different body poses) that can be selected for mapping.

Consider an object process O having probability

$$\pi(\mathbf{o}) = \prod_i \pi(\mathbf{o}_i) = \prod_i \pi(p_i)\pi(w_i, h_i, \theta_i|p_i)\pi(s_i), \quad (1)$$

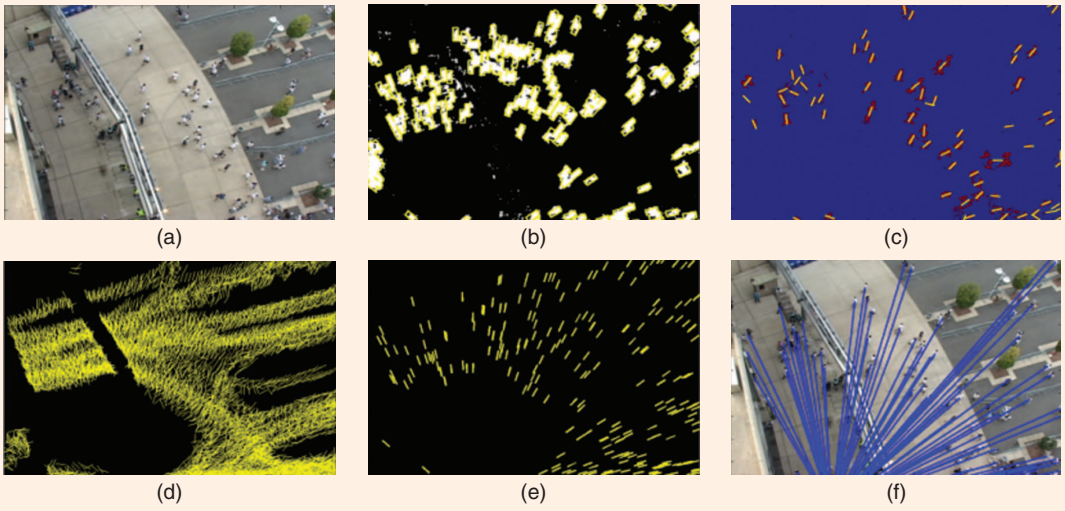
where $\pi(p_i)$ is a homogeneous Poisson point process that governs the spatial distribution of people in the scene and $\pi(w_i, h_i, \theta_i|p_i)$ is a conditional mark process for extrinsic shapes, e.g., rectangles representing the shape and orientation of a two-dimensional (2-D) bounding box, conditioned on spatial location. A Poisson process is chosen for its complete spatial randomness property [1] that implies independence between disjoint regions within the observation image. We do not want to impose a prior that would induce spatial patterns such as clustering effects, since we want to generalize our crowd model to different video sequences. The extrinsic shape process encodes our prior knowledge of strong correlations between the size and orientation of projected objects and their 2-D image locations, in views taken by a static camera. In addition to extrinsic shape, each person is also associated with an intrinsic shape to model different pedestrian poses, and $\pi(s_i)$ is a uniform distribution over the shape prototype index set S.

ESTIMATING EXTRINSIC SHAPES

We represent the conditional mark process $\pi(w_i, h_i, \theta_i|p_i)$ by independent Gaussian distributions describing the expected width, height, and orientation of a pedestrian bounding box centered at image location p_i. The means of these Gaussian distributions are automatically estimated from a small sample of the sequence where the crowd density is low. Inferring camera calibration parameters from watching people in the scene has also been considered in [3] and [4].

Since we know pedestrians will be oriented vertically, the vertical vanishing point of the scene completely determines the 2-D image orientation of a person at any location. Figure 1 illustrates computation of the vertical vanishing point for a sample sequence. Foreground masks are computed for each frame via background subtraction. Blobs are found by connected components, followed by ellipse fitting to compute their center of mass and second moments. In Figure 1(e), we repeat the process of extracting major axis orientation of blobs for all frames in a short sequence of video. Some of the axes

1053-5888/10/$26.00©2010IEEE



[FIG1] The image orientation of a standing person at any image location is determined by automatically estimating the vertical vanishing point of the scene from video of walking pedestrians: (a) original image, (b) foreground blobs, (c) blob orientations from one frame, (d) blob orientations from many frames, (e) inliers found by RANSAC, and (f) vertical vanishing point.

represent vertical orientation of individuals who are found as a single blob; however, many others are outliers representing the orientation of multiperson blobs, fragmented blobs, or blobs whose second moments are corrupted away from vertical by arms and legs extending out from the person. To find the vertical vanishing point, we assume that the inlier axes will converge to a vanishing point and use random sample consensus (RANSAC) to find the intersection point voted for by the most axes. We see that the computed vanishing point correctly captures the change in image orientation of people at different parts of this scene. The orientation of a blob centered at any pixel in the image can now be computed and stored in a lookup table representing the mean of a Gaussian distribution on orientation.

Given blob orientation at each point in the image, blob height and width are computed with respect to that orientation. A reasonable first-order model of many scenes assumes that people are walking or standing on a planar ground surface. This planarity assumption regularizes the computation of size by constraining the relative depth of people in the scene as a smooth function of image location. Similar to the computation of orientation, height, and width are also computed from observations of walking pedestrians in a training sequence by robustly fitting parametric functions for height and width and forming lookup tables representing their Gaussian mean values at any image pixel.

LEARNING INTRINSIC SHAPES

Simple geometric shapes such as rectangles or ellipses are only a coarse approximation to the shape of people we want to count. In this section, we learn the parameters of a mark process that well approximates the appearance of foreground shapes.

Rather than treating all pixels in a rotated and scaled bounding box as foreground, we consider a "soft" segmentation of shape, representing the probability of each pixel being foreground. We use a mixture of Bernoulli distributions to model learned shape prototypes as rectangular patches of spatially varying $\mu(x_i)$ values, one per pixel, learned from a training set of observed foreground masks. The μ values are high in areas of the rectangle that often contain foreground pixels, and low in places that often contain background, as visualized in the grayscale shape images in Figure 2. The mixture model can represent a varied and realistic set of shape prototypes, resulting in more accurate foreground fitting.

To learn the shape prototypes from a training video sequence, we first select a random subset of frames labeled with ground truth bounding boxes, run background subtraction to get binary masks, which yields a set of binary shape patterns, then scale each shape to a standard size. Denote $\mathbf{X} = \{\mathbf{x}_i, \ldots, \mathbf{x}_N\}$ as the collection of N training shape patterns, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^T$ (D being the size of the shape pattern) and each x_{ij} is a binary variable. We model $\mathbf{X}$ by a mixture of Bernoulli distributions. Our choice of Bernoulli distribution as the component distribution for the mixture model is motivated by its success in recognizing the shape of handwritten digits [5]. Formally, the mixture model is defined as

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(x|\boldsymbol{\mu}_k), \qquad (2)$$

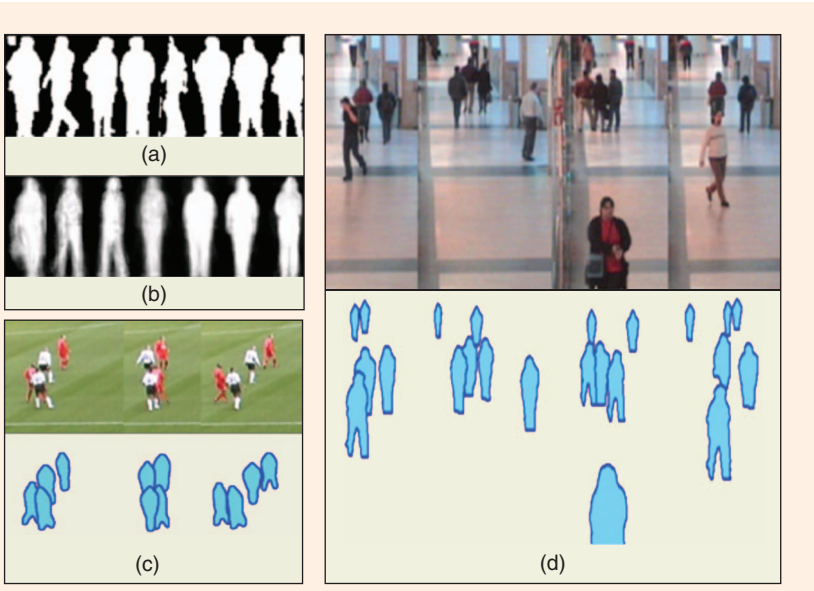
where K is the number of mixture components, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ are the Bernoulli mean parameters, each of which is itself a vector $\boldsymbol{\mu}_k = (\mu_1, \ldots, \mu_D)^T$, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ are the component mixing weights, and $p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{d=1}^{D} \mu_d^{x_d}(1-\mu_d)^{(1-x_d)}$ is one component Bernoulli distribution. We extend the above classic mixture model to a weighted Bernoulli mixture, motivated by the observation that certain pixels vary more across different shapes than other pixels.

For example, the boundary pixels of the body shape usually have larger variance than the background pixels or pixels surrounding the center of mass. It is therefore advantageous to make the model spend more effort explaining the higher-variance parts of the shape so that we can get a better shape class model with more distinctive components. For this purpose, we introduce pixel-wise weights $\mathbf{v} = (v_1, \ldots, v_D)^T$ that are estimated variance at each pixel across all the training patterns. Hence, $p(\mathrm{x}|\boldsymbol{\mu}_k)$ can be rewritten as

$$p(\mathrm{x}|\boldsymbol{\mu}_k) = \prod_{d=1}^{D} \mu_d^{x_d v_d} (1-\mu_d)^{(1-x_d)v_d}. \quad (3)$$

The complete derivation of the above equation is provided on our project Web page, http://vision.cse.psu.edu/projects/mpp/mpp.html, but the intuition is simple: we can treat the weight as a replication factor; the higher the weights, the more important the pixels and the more times they get duplicated in the sample.

One typical difficulty with using mixture models is how to determine the number of components. We automatically determine the number of components K by imposing a Dirichlet prior over the mixing weights $p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$. By setting $\alpha_k \simeq 0$, we have a broad prior that squashes some of the mixing weights to zero, i.e., the Bayesian model automatically balances quality of fit to the data and the complexity of the model. Thresholding on α_k automatically determines the number of intrinsic shapes learned. In our experiments, we set α_k to be a small positive number. All model parameters are estimated by the expectation-maximization algorithm [2]. The appeal of the weighted Bernoulli shape mixture model is that the parameter estimation is very efficient, yet the model itself is flexible enough to be generalized to encode different shapes (Figure 2).



[FIG2] **Intrinsic shape classes are modeled by a mixture of Bernoulli distributions, learned from binary image patches extracted in a training sequence. Detection results show that the shape covering accurately characterizes different pedestrian poses (e.g., legs together/apart). (a) Training samples, (b) automatically learned shapes, (c) detecting soccer players, and (d) detecting people in a shopping mall.**

INFERENCE BY RJMCMC

The observed data is a foreground mask, assuming the foreground is formed by pedestrians in the scene. A generative explanation of this foreground mask involves selecting the appropriate intrinsic shape prototypes and then translating, rotating, and scaling them into the image to cover the foreground pixels as well as possible. To compute the goodness of fit of a proposed configuration of shapes to the data, we adopt a likelihood function similar to previous works [6]–[8]. First the configuration is mapped into a soft label image. Let x_i be the values in the label image and y_i the binary values in an observed foreground mask. The pixel values in the label image are continuous random variables ranging from $[0, 1]$, parameterized by the mean of the Berouli distribution $p(y_i|x_i)$. Assuming conditional independence among the pixels, the joint log likelihood function can be written as

$$\begin{aligned} \log \mathcal{L}(\mathrm{Y}|\mathrm{X}) &= \log \prod_{i=1}^{N} p(y_i|x_i) \\ &= \sum_{i=1}^{N} (y_i \log x_i + (1-y_i) \\ &\quad \times \log(1-x_i)). \end{aligned} \quad (4)$$

This likelihood function as written does not discourage configurations having multiple overlapping shapes that claim almost the same set of foreground pixels. To avoid this, we implement a simple scheme where the number of overlapping pixels is multiplied by a nonnegative factor ρ to form a penalty term subtracted from the log likelihood function.

The likelihood function and the MPP prior (1) combine to form a posterior that measures how well the observed foreground mask can be described as a noisy instantiation of our statistical crowd model. Pedestrian detection and counting then becomes the problem of estimating the maximum a posteriori (MAP) configuration. We use RJMCMC sampling [9], [10] to perform Bayesian inference of the best crowd configuration. RJMCMC is an iterative sampling procedure that proposes either a local update to a current configuration or a reversible jump between configurations of differing dimensions, and then decides stochastically whether or not to accept the new configuration based on the value of a ratio

$$a(\mathbf{o}, \mathbf{o}') = \min\left(1, \frac{p(\mathbf{o}')q(\mathbf{o}', \mathbf{o})}{p(\mathbf{o})\, q(\mathbf{o}, \mathbf{o}')}\right),$$

where $\mathbf{o}$ and $\mathbf{o}'$ are the current and proposed configurations, $p(\cdot)$ is the posterior distribution evaluated for a given configuration, and $q(a, b)$ is the probability of proposing a transition from a

to b. We use a simple RJMCMC sampler composed of birth, death, and update proposals [8], [11]. Figure 3 illustrates how different proposals switch between configurations. Each of the proposals is described briefly as follows:

1) Birth: A point and mark are proposed and added to the current configuration. We sample the point location with a data-driven proposal based on the foreground mask. Width, height, and orientation of the rectangular mark are sampled from the conditional mark process, represented as Gaussian distributions indexed by spatial point location. An intrinsic Bernoulli shape is chosen uniformly at random (u.a.r) from the set of learned shape prototypes. The reverse move of birth is death.

2) Death: The death proposal chooses one rectangle at random and removes it from the configuration. The reverse move is birth.

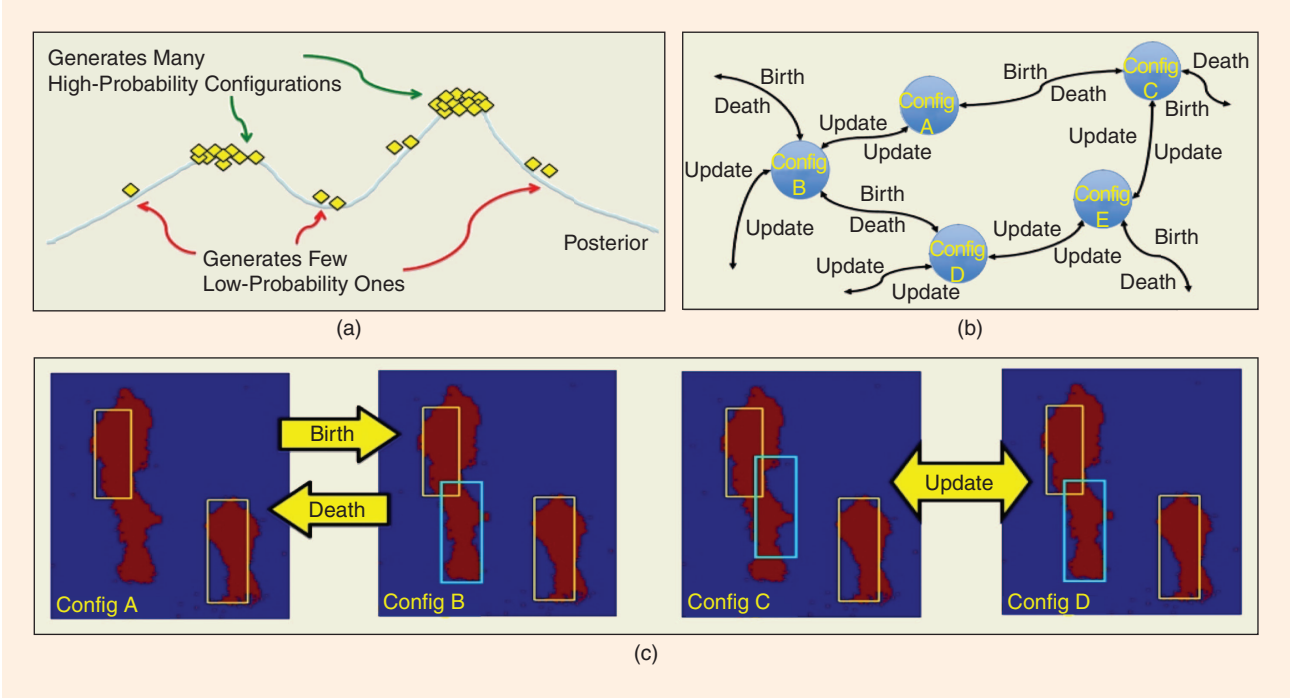
3) Update: One rectangle from the configuration is chosen at random and either its location or mark parameters are modified. Modification of location is done as a random walk of the shape center. Modification of the mark is done in two parts: either the mark width, height, and orientation are updated by sampling from the conditional mark process associated with the current location, or the intrinsic Bernoulli shape is updated u.a.r from the shape prototype set. The update proposal is its own reverse move.

Starting with an empty configuration as the initial state, the RJMCMC procedure is iterated between 500 and 3,000 times, with the larger number of iterations being needed when there are more people in the scene. The move probability for birth, death, and update proposals is set to be 0.4, 0.2, and 0.4, respectively. During iterations, the configuration with the highest observed posterior probability is saved, and at termination that configuration is output as an estimate of the MAP solution.

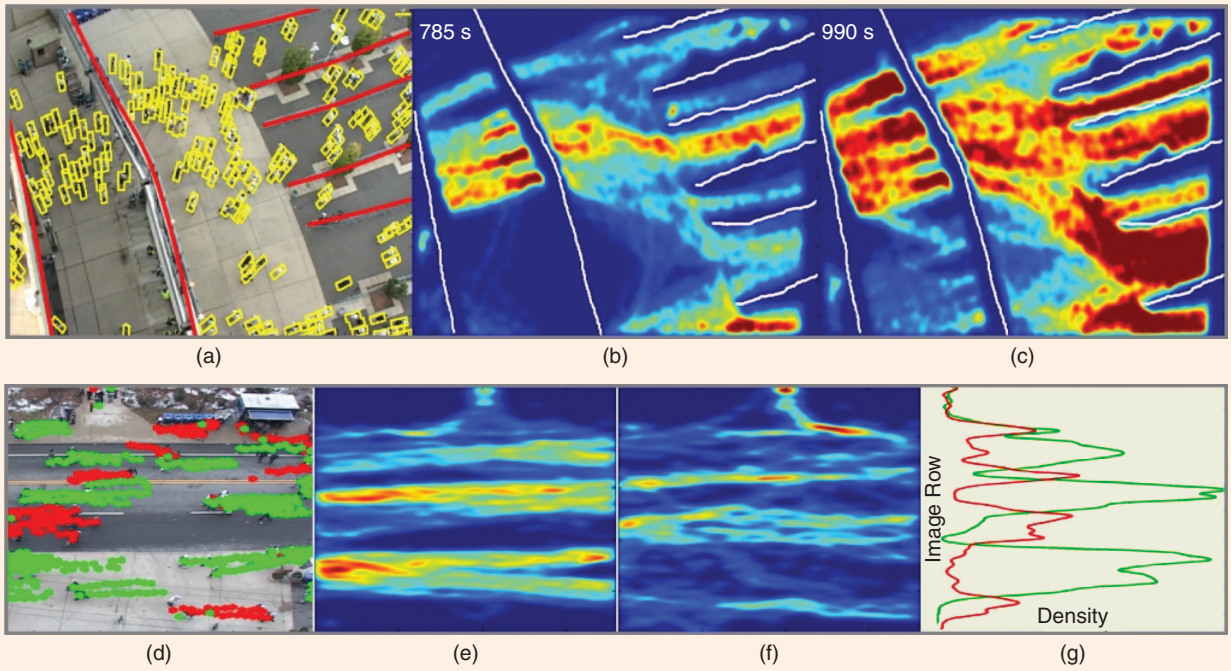
CROWD DENSITY ANALYSIS

Our MPP model with shape appearance is capable of detecting people in crowds with various densities, from different viewing angles, and in indoor and outdoor scenes. Figure 2 shows sample detection results on two benchmark sequences: the EU CAVIAR data set (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/) and the VSPETS soccer sequence (http://www.cvg.cs.rdg.ac.uk/VSPETS/vspets-db.html). The CAVIAR data set contains sequences of people walking in a shopping mall. People are observed at a large range of image sizes as they travel down the hallway. The soccer sequence was captured in an outdoor football field. Players have roughly the same size throughout the image but often occlude each other as they run back and forth. Our model captures different human body shapes well, as shown in Figure 2.

We also analyzed crowd density outside a football stadium, using data collected with a Sony DCR VX2000 digital video camcorder mounted on the stadium. The viewpoint is highly elevated and the image size of each person is relatively small, thus we use



[FIG3] **Parts (a) and (b) illustrate the principles of MCMC search. The MCMC sampler automatically spends more effort generating high probability samples from a configuration space by exploring an implicit local neighborhood of configurations generated from a set of proposal moves. Part (c) shows how the different proposals (birth, death, and update) change a current configuration into a new hypothesis.**



[FIG4] Two different crowd density analyses are shown in (a)–(c) and (d)–(g). (a) Crowd detection overlaid on an image of people leaving a football stadium, with red lines delineating the major lanes of egress. (b) and (c) show kernel density estimates of crowd density over spatio-temporal windows centered at two different times. (d) Centroids of the detections overlaid on an image of pedestrians walking on a road outside the stadium prior to the game. Detections moving left are colored in green and rightward detections are colored in red. (e) Crowd density of leftward traffic. (f) Crowd density of rightward traffic. (g) Confirmation that individuals in the crowd unconsciously form anticorrelated lanes of traffic, the well-known "fingering effect."

simple rectangular shapes instead of Bernoulli shapes for detection. Figure 4 shows sample detections and crowd density estimates. We analyze crowd densities at each image location by kernel density estimation from detections within a spatio-temporal window. The visualized density maps at two different times [Figure 4(b) and (c)] reflect the change in traffic flow patterns due to an increased number of people leaving after the game ended. The crowd density analysis in another sequence [Figure 4(d)–(g)] looking down at a road outside the stadium shows an interesting crowd behavior pattern: people moving in opposite directions, left and right, tend to unconsciously form lanes of traffic to avoid collisions, helping the entire crowd move more smoothly. As a result, the leftward and rightward crowd densities exhibit an anticorrelated pattern [Figure 4(g)], known as the "fingering effect" within the study of crowd dynamics.

SUMMARY AND ONGOING WORK

We have presented an MPP model for detecting people in crowds, with a mark process parameterized by extrinsic appearance (geometry) and intrinsic appearance (shape and posture), learned separately from training sequences. The optimal crowd configuration is estimated by RJMCMC sampling methods and used for crowd scene analysis under different crowd densities and environmental situations. Our current approach analyzes each frame independently, thus is suitable for counting people in sequences with very low frame rate or for automatic initialization of a multitarget tracker. However, it is also possible to incorporate temporal coherence and to integrate detection and tracking within the same sampling framework by augmenting the state space with temporal variables, similar to [12]. The Bayesian formulation for learning weighted Bernoulli shape masks is very flexible, and can be used to model a variety of shapes within the MPP framework. In follow-up work, we also plan to consider different object classes (cars, bikes, and pedestrians) using the same model framework, with potential applications to activity monitoring at road intersections.

ACKNOWLEDGMENT

This work was partially funded by the NSF under grants IIS-0729363 and IIS-0535324.

AUTHORS

Weina Ge (ge@cse.psu.edu) is a Ph.D. candidate in the Computer Science and Engineering Department at Penn State University, specializing in crowd image analysis.

Robert T. Collins (rcollins@cse.psu.edu) is an associate professor in the Computer Science and Engineering Department at Penn State University, where he codirects the Laboratory for Perception, Action, and Cognition and conducts research into video scene understanding.

(continued on page 123)

Needless to say, I am delighted by these technology developments, not only because of their convenience and expansive educational potential, but because all of it, the cameras, displays, broadcast system, video courseware, and video recordings, are examples of byproducts of digital video processing research. It is very satisfying to lecture using the technology I am teaching.

Looking ahead, there are more exciting developments, not the least of which is 3-D. The movie *Avatar* has raised public awareness of the amazing experiences to be found in cinematic 3-D video. More importantly, 3-D technology is going to significantly penetrate the broader consumer market soon—3-D televisions are already commercially available, and glasses-free auto-stereoscopic displays will soon be good enough (and cheap enough) for the home audience as well. These displays will also be found on handheld devices. We aren't to the point of Princess Leia calling for "Obi-Wan Kenobi" via holo-projection, but we aren't far either.

These 3-D technologies will be available in the classroom as well. Before long, 3-D classroom displays will not be uncommon, and given the exposure and commercial drive in this direction, 3-D video instruction (meaning 3-D topics) and 3-D instruction techniques (meaning teaching in 3-D) are obvious developments to look forward to.

I hope that I have been able to express the enthusiasm and joy I find in teaching digital video processing. As I approach 30 years as a professor, there are many things that I do not look forward to every day, but one thing I always anticipate is lecturing on digital video.

RECOMMENDATION 6

Toss the chalk and whiteboard marker! Use modern video acquisition, communication and interactive display technology to teach digital video processing!

ACKNOWLEDGMENTS

The SIVA courseware was largely created by three students: Umesh Rajashekar, George Panayi, and Frank Baumgartner. Panayi wrote nearly all of the still image SIVA demos under my direction, while Baumgartner wrote the video SIVA demos. Rajashekar wrote all of the audio and one-dimensional demos which are written in MATLAB. He has also maintained the SIVA system (long after leaving UT at Austin), created the SIVA Web site, and coauthored our papers and reports on SIVA.

NI, Inc., located here in Austin, was unfailingly supportive in the development of the SIVA system. They were kind enough to fund the efforts of Panayi and Baumgartner (both NI engineers as well as UT students), and have taken an active role in maintaining SIVA through every Labview upgrade. More recently, several NI engineers (Nate Holmes, Matthew Slaughter, Carleton Heard, and Nathan McKimpson) assisted Rajashekar and me in creating a standalone (compiled) version of the SIVA image processing demos, complete with global user interface (from which all demos are easily accessed), for inclusion on a CD-ROM with my recent book [12]. NI manager Dinesh Nair assisted us in overseeing these efforts and in Chapter 2 of [12] helped us explain the system and the use of Labview.

REFERENCES

[1] D. S. Teyhen, T. W. Flynn, A. C. Bovik, and L. D. Abraham, "Digital fluoroscopic video assessment of sagittal plane lumbar spine flexion," *Spine,* vol. 13, no. 1, pp. E406–E413, Jan. 2005.

[2] M. Tur, K. C. Chin, and J. W. Goodman, "When is speckle noise multiplicative?," *Appl. Opt.*, vol. 21, no. 7, pp. 1157–1159, 1982.

[3] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr., "Non-linear filtering of multiplied and convolved signals," *Proc. IEEE,* vol. 56, no. 8, pp. 1264–1291, Aug. 1968.

[4] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B.,* vol. 207, no. 1167, pp. 187–217, Feb. 1980.

[5] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[6] B. Julesz, *Foundations of Cyclopean Perception*. Chicago, IL: Univ. of Chicago Press, 1971.

[7] K. Seshadrinathan and A. C. Bovik, "Motion-tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Processing,* vol. 19, no. 2, pp. 335–350, Feb. 2010.

[8] S. E. Palmer, *Vision Science: From Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[9] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust, "Do we know what the early visual system does?," *J. Neurosci.,* vol. 25, no. 46, pp. 10577–10597, Nov. 2005.

[10] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using Matlab.* Gatesmark, 2009.

[11] U. Rajashekar, G. Panayi, F. P. Baumgartner, and A. C. Bovik, "The SIVA demonstration gallery for signal, image, and video processing education," *IEEE Trans. Educ.,* vol. 45, no. 4, pp. 323–335, Nov. 2002.

[12] A. C. Bovik, *The Essential Guide to Image Processing*. New York: Elsevier Academic Press, 2009.

[SP]

[applications CORNER] continued from page 111

REFERENCES

[1] M. van Lieshout, *Markov Point Processes and Their Applications*. London: Imperial College Press, 2000.

[2] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2913–2920.

[3] N. Krahnstoever and P. Mendonca, "Bayesian autocalibration for surveillance," in *Proc. IEEE Int. Conf. Computer Vision,* Oct. 2005, pp. 1858–1865.

[4] D. Rother, K. A. Patwardhan, and G. Sapiro, "What can casual walkers tell us about a 3d scene?," in *Proc. IEEE Int. Conf. Computer Vision,* Oct. 2007, pp. 1–8.

[5] A. J. Baddeley and M. vanLieshout, "Stochastic geometry models in high-level vision," in *Statistics and Images,* vol. 1, K. V. Mardia and G. Kanji, Eds. Nashville, TN: Abingdon, 1993, pp. 231–256.

[6] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 30, no. 2, pp. 267–282, 2008.

[7] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2003, pp. 459–466.

[8] P. Green, "MCMC in image analysis," in *Markov Chain Monte Carlo in Practice,* W. R. Gilks, S. Richardson, and D. Spiegelhalter, Eds. London, U.K.: Chapman & Hall, 1995, pp. 381–400.

[9] H. Rue and M. Hurn, "Bayesian object identification," *Biometrika,* vol. 86, no. 3, pp. 649–660, 1999.

[10] X. D. M. Ortner and J. Zerubia, "A marked point process of rectangles and segments for automatic analysis of digital elevation models," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 30, no. 1, pp. 105–119, 2008.

[11] X. D. M. Ortner and J. Zerubia, "A marked point process of rectangles and segments for automatic nalysis of digital elevation models," *IEEE TPAMI,* vol. 30, no. 1, pp. 105–119, 2008.

[12] K. Otsuka and N. Mukawa, "Multiview occlusion analysis for tracking densely populated objects based on 2-D visual angles," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004, pp. 90–97.

[SP]