Video-Based Crowd Synthesis

Matthew Flagg, *Member*, *IEEE*, and James M. Rehg, *Member*, *IEEE*

Abstract—As a controllable medium, video-realistic crowds are important for creating the illusion of a populated reality in special effects, games, and architectural visualization. While recent progress in simulation and motion captured-based techniques for crowd synthesis has focused on natural macroscale behavior, this paper addresses the complementary problem of synthesizing crowds with realistic microscale behavior and appearance. Example-based synthesis methods such as video textures are an appealing alternative to conventional model-based methods, but current techniques are unable to represent and satisfy constraints between video sprites and the scene. This paper describes how to synthesize crowds by segmenting pedestrians from input videos of natural crowds and optimally placing them into an output video while satisfying environmental constraints imposed by the scene. We introduce *crowd tubes*, a representation of video objects designed to compose a crowd of video billboards while avoiding collisions between static and dynamic obstacles. The approach consists of representing crowd tube samples and constraint violations with a conflict graph. The maximal independent set yields a dense constraint-satisyfing crowd composition. We present a prototype system for the capture, analysis, synthesis, and control of video-based crowds. Several results demonstrate the system's ability to generate videos of crowds which exhibit a variety of natural behaviors.

Index Terms—Crowd synthesis, video-based rendering, video object layout, crowd constraint optimization

1 Introduction

As a controllable medium, video-realistic crowds are important for creating the illusion of a populated reality in special effects, games, and architectural visualization. Plausible crowd imagery should depict a layout-constrained scene with people entering and exiting at specific locations and moving through the environment while responding to dynamic obstacles. Individuals comprising a crowd should also exhibit a variety of behaviors to appear realistic. For example, pedestrians may move in a hurried, goal-directed manner, or meander while curiously inspecting other people and points of interest. In current practice, crowd artists animate and render crowd effects via a tedious model-based graphics pipeline.

Model-based crowd animation is made possible at a large scale via simulation, by crafting rules for individual agents, thus freeing the artist from having to painstakingly place every keyframe for each control rig. Most research progress has focused on simulating large numbers of agents that exhibit macroscale behaviors like lane and vortex forming, which are well suited for background or distant crowds [1], [2], [3]. Unfortunately, simulated mid-ground and foreground crowds require an expensive and complex production pipeline to convey realism in cloth, body, hair, and facial motion, especially when approaching the level of detail exhibited by up-close hero characters. Fig. 1 illustrates several examples of complex behavior which would be difficult to achieve with current model-based crowd synthesis techniques.

• *The authors are with the Georgia Institute of Technology, College of Computing Building, 801 Atlantic Drive, Atlanta, GA 30332. E-mail: {mflagg, rehg}@cc.gatech.edu.*

Manuscript received 3 Sept. 2011; revised 4 June 2012; accepted 22 Nov. 2012; published online 29 Nov. 2012.
Recommended for acceptance by D. Schmalstieg.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2011-09-0215.
Digital Object Identifier no. 10.1109/TVCG.2012.317.

When carefully planned, a crowd artist or director may successfully compose their desired crowd by filming and building it up person-by-person or one block at a time. Unfortunately, crowd designers rarely have the resources or ability to edit the layout by gathering blocks of actors for additional takes. For crowds that exhibit detailed behaviors at the individual level, such as tousling one's hair just as another walks by, it is difficult to satisfy behavioral space-time constraints in addition to collision avoidance, constrained movement from entry to exit, desired levels of crowd density, and clone separation to preserve the perception of crowd variety.

To explore realistic crowd synthesis, we adopt the philosophy of example-based rendering: by directly copying pixels from input imagery of real crowds to output images of synthetic crowds, realism may be attained. This paper addresses the problem of photo-realistically depicting a variety of natural crowd imagery by composing video clips of individual and group behavior subject to environmental constraints. Our method is capable of synthesizing video-based scenes that display a variety of contextually plausible behaviors because they are captured from the target output environment. For example, one would expect to see more sight-seeing behaviors in a crowd at the Embarcadero of San Francisco than a major thoroughfare at a college campus between classes.

In contrast with model-based crowd simulation, our approach to crowd synthesis does not involve parameter estimation and tuning to achieve aggregate dynamics. We assume that a plausible crowd may be built up from individual and group behaviors and can support the creation of aggregate effects, such as lane forming, by capturing and reusing natural examples of such macroscale behaviors. In building our system, we identified and addressed several challenges to video-based crowd synthesis.

Fig. 1. Video-based crowd. Top row: video clips of natural individual and group behaviors serve as input. Bottom: a still from a crowd output video produced using our system, which inherently exhibits realistic behaviors.

1.1 Challenges

Video-based crowd synthesis must address two categories of technical challenges: 1) Synthesis is a constrained layout problem and constraints must be represented and enforced while preserving variety, and 2) video-based crowds must be segmented and reused while keeping video objects intact and smooth in motion.

Layout constraints. A video of a crowded scene may exhibit large variations in the density, appearance, and motion of its pedestrians. The presence of such variations is important to crowd plausibility yet is difficult to achieve due to an abundance of layout constraints. The process of copying input pixels to the output must satisfy environmental constraints imposed by the scene. For example, crowds should not float in mid-air above the ground planes, walk in traffic, or through walls of buildings. A crowded scene exhibits static and dynamic obstacles such as mailboxes and pedestrians that further constrain the layout of animated video clips.

Video segmentation and reanimation. In addition to satisfying layout constraints, a video-based crowd exhibiting good crowd variety requires the ability to clip out and reuse recorded clips of crowd behavior in a perceptually convincing manner. Straightforward techniques for video segmentation, such as background subtraction, are usually insufficient as they do not enforce temporal coherence to prevent pieces of foreground, such as heads and arms of pedestrians, from noticeably scintillating during playback as they oscillate between inclusion and exclusion from the video object. Following segmentation, crowd synthesis must copy and place segmented clips of behavior back into the scene. Current techniques for video resynthesis, such as video textures and video sprites [4], are incapable of identifying transition points in articulated pedestrian video.

1.2 Summary of Approach

In this paper, we focus on layout constraint satisfaction, the first category of technical challenge. To address the second category, we employed straightforward techniques that merit further attention in future work. Commercial video object segmentation was employed to clip out examples of behavior from a fixed view. Transition frames were manually identified in each input clip for the purpose of video duration extension via concatenation at transitions. However, it is possible to generate a greater variety of video-based pedestrian animations over a motion graph [5]. The video results illustrate the appearance of crowds constructed using crowd tubes before and after satisfying constraints, thus demonstrating the perceptual effect of increasing crowd density by relaxing collision and variety constraints. Resulting animations are inherently limited by the number of captured behaviors and viewpoints. Therefore, the crowd solution space with our video-based approach is more constrained than traditional simulation. Limitations are described in detail in Section 8.

Evaluation. Using a prototype system for video-based crowd synthesis, we evaluated our approach in two ways: 1) demonstration of output crowd videos from four scenes including sailboats and pedestrians,[1] and 2) a user evaluation. Three users planned their vision of a crowd given a palette of captured crowd behaviors. After creating their plans in the form of hand-drawn sketches, they used the prototype system to execute their vision.

Contributions. This paper presents three contributions:

- Crowd tubes, a novel representation of the visual elements of a crowd for synthesizing crowd video including individuals, groups of individuals, and inanimate objects, coupled with temporally varying Three-dimensional shape proxies, which enable placement of crowd elements in a Three-dimensional scene while satisfying constraints from the scene.
- Formulation of a constraint satisfaction problem based on crowd tubes, which enables the synthesis of video-based crowds with properties such as collision avoidance and spatial separation of clones that are necessary for realism.
- The first experimental results for video-based crowd synthesis from video, which illustrates the effectiveness of crowds before and after satisfying constraints on crowd tubes

These results extend previous techniques in computer graphics for crowd synthesis and demonstrate the feasibility of video-based crowd composition.

2 Background

Traditional crowd synthesis in computer graphics employs techniques for modeling crowd behavior both procedurally and from data. In contrast with our method, these techniques output joint angles and trajectories instead of pixels. The film industry frequently uses Massive [6], [7], a popular agent-based crowd simulation system in which each agent plans their own motion individually. Sophisticated agent-based methods model behavioral dynamics, or how to respond to stimulus, cognitive aspects such as terrain knowledge and learning, and pedestrian visibility for path planning [8], [9]. Many researchers have also tried

1. A video of several crowd results before and after constraint satisfaction is available here: http://youtu.be/ygTsoninp-Q.

simpler agent-based models that avoid cognitive modeling [10], [11], [12], [13], [14], [15]. In this section, we review particle and image-based crowds as background for the proposed video-based technique.

2.1 Particle-Based Crowds

Early work on flocking and herding treated crowd simulation as an elaboration of particle systems, with the boid flock model [16] being a seminal example. Particles in the boid flock model are represented with oriented objects, such as birds, which exhibit complex behaviors governed by internal bird state and the external flock state. This approach was an alternative to scripting the paths of each flock component individually. The boid flocking model is based on the following three heuristic rules in order of decreasing precedence: 1) collision avoidance, 2) velocity matching: match velocity with nearby flockmates, and 3) flock centering.

In contrast, the proposed approach manipulates a set of video objects and their Two-dimensional ground plane trajectories to animate a video-based crowd. Thus, our crowd model is closer to the alternative method to Reynold's boid that consists of scripting the paths of each crowd component individually. Our trajectories are data driven, however. A fixed family of trajectories is the set of possible crowd tubes that may be plausibly generated from a single crowd tube. A family of trajectories is generated via translation about the ground plane and temporal resequencing.

It is not clear how much shape change a crowd tube could undergo before damaging plausibility. Consider the appearance of a pedestrian recorded while walking in a direction that is fronto-parallel to the camera's image plane. How can we reanimate this example to walk away from the camera in the depth direction? The back is not visible and the motion of the feet is incorrect. Resulting footskate artifacts could potentially be hidden in a dense crowd, but trajectories that disagree with body orientation seen in the head and upper torso may be noticeable. For example, a forward-facing pedestrian moving sideways may give away its synthetic nature, even if the head is solely visible.

Particle-based crowds rest upon the assumption that control rigs can be animated to move along arbitrarily shaped ground tracks. Potential fields used in [1] have been applied to achieve effects such as lane forming and bottlenecking. Massive animates lower level control rigs by blending motion capture data. A related motion capture-driven technique for synthesis of crowds is the method of [17]. A traditional motion graph [18], [19] is used with probabilistic maps to build a crowd of goal-directed individuals one at a time in a greedy fashion. The method makes potentially severe updates to joint angles to satisfy hard space-time constraints.

Crowd patches [20], an extension of motion patches [21], efficiently simulates large crowds for real-time applications by tiling blocks of precomputed local crowd simulation. A crowd patch is space-time block of precomputed trajectories that are cyclic over a constant period. By interconnecting neighboring patches with compatible trajectory entry and exit points, animated objects appear to seamlessly cross the limits of a patch to a neighbor during looping playback. However, the technique would require a tedious planning and capture process to design scene-dependent crowd patches, direct pedestrians to carefully satisfy patch boundary conditions at the right point in space and time, and synthesize transitions needed in the beginning and end of each period.

In contrast to constructing tile-like examples of crowds for a given spatial area, our approach more closely resembles the example-based methods of [22], [23], [24] to simulate crowds. Lerner et al. use a set of trajectories extracted manually from crowd video to control autonomous agents. Agent trajectories are incrementally synthesized by considering spatiotemporal relationships with nearby agents and searching for similar scenarios in a trajectory database. They employ locally weighted linear regression to model an agent's speed and moving direction as a function of the motion of nearby agents, its own motion and environment features. The dynamic model is applied to animate traditional model-based avatars. Our system takes a more extreme example-based approach to attain photorealism by reusing captured trajectories and pixels (represented by crowd tubes) with modifications limited to translation and rearrangement in time.

2.2 Image-Based Crowds

A number of researchers have exploited image-based proxy representations of pedestrians for rendering crowds of thousands efficiently [25], [26]. These methods enable fast drawing by instancing billboards mapped with textures obtained by rendering more detailed Three-dimensional models. Our approach trades such rendering efficiency for visual quality by instancing billboards texture mapped with real video.

Lalonde et al. describe a closely related system for inserting new objects into existing photographs by sampling from a large image-based object library [27]. Their key idea is to search a database of over 13,000 objects for examples with lighting that is consistent with the scene. Like our system, theirs has a need to estimate the ground plane for perspectively correct composition. Another similarity to our problem is the need for high-quality segmentation and matting. Our goal for video-based crowd composition faces layout challenges which they do not address. Their system's ability to produce plausible photographs gives us inspiration the crowd video domain.

3 Problem Statement

Given an input video of a sparsely crowded scene recorded by a camera with a fixed center of projection, synthesize a video of the same scene populated with instances of recorded video objects of artist-controlled duration, density, and behavior. Assume the input contains the following video segments obtained by panning, tilting, and zooming as captured by a tripod mounted camera.

- V, a zoomed out view of the scene to be used as a background plate during crowd synthesis.
- V_i, a set of zoomed in views indexed by variable i, which contain fixed view footage of individuals and groups of individuals targeted by the recorder. V_i serves as a "palette" of natural behaviors that comprise the video-based crowd medium.

A plausible crowd video must satisfy constraints on the appearance and motion of crowd elements exhibited in V_i.

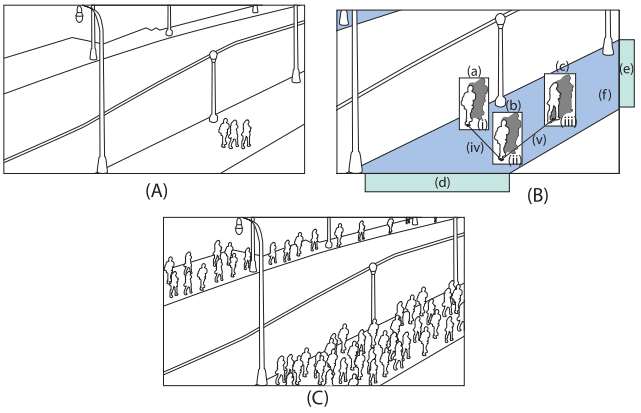


Fig. 2. Input, constraints, and output. A) Input: sparse crowd video in the target scene. B) Constraints: segmented video clips are composed in the output video subject to a minimal separation distance (v) between ground locations (ii, iii) for collision avoidance, a minimal separation distance (iv) between ground plane locations (i, ii) for clones, entry and exit regions of interest (d and e) and pedestrian traversability regions (f). Transition visibility is approximated by the distance between a transition frame of one crowd tube and a non-transition frame of another crowd tube (e.g., distance (v) when (c) is transitioning and (b) is not). C) Output: A crowd video that satisfies layout constraints.

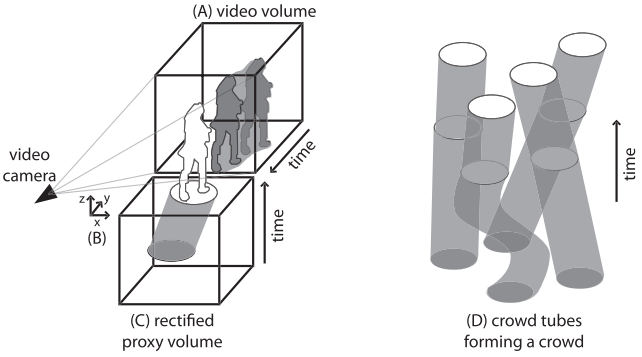


Fig. 3. Rectified proxy volume. Segmented video clips require Three-dimensional scene information to satisfy constraints such as collision avoidance. A) The video volume of a segmented pedestrian carves out a tube-like shape. B) The pedestrian's ground track in the input video is converted to world coordinates. C) An ellipse planted in the ground plane for each captured frame serves as a proxy for the pedestrian's ground occupancy. Over time, this produces a proxy volume corresponding to the video volume. By rectifying each frame of the video volume to that of a top-down view, a ground plane-over-time representation enables constraint enforcement (collision avoidance, spatial separation of clones, and so on). D) A collection of nonintersecting crowd tubes yield a crowd. Ellipses represent transitions in captured footage.

- Elements must rest on the ground plane in the scene and exhibit correct perspective.
- Elements can only occupy traversable sections of the ground plane.
- Elements must enter and exit from semantically correct areas of the scene (i.e., they must not appear and disappear in free space).
- No pair of elements should collide (occupy the same space on the ground plane).
- No pair of elements cloned from the same V_i should be in close proximity. This perceptual constraint is motivated by recent findings from a study of the perception of crowd variety [28].

Fig. 2 illustrates the problem setup including an input view, constraints necessary for plausibility, and an example output view.

4 Satisfying Crowd Constraints

We approach the problem by treating it as a rigid trajectory layout problem. Each V_i is segmented and its ground plane track is used to animate an alpha-matted video texture billboard. A video texture billboard or video sprite is a camera-facing Three-dimensional plane texture mapped with a video texture [4]. The ground plane track is automatically computed from a segmented V_i by raycasting the ground contact point through a calibrated Three-dimensional ground plane. Section 5.3 describes the calibration steps in detail.

A crowd artist interactively specifies unary constraints on *crowd tubes*, our novel representation of crowd behaviors as rigid trajectory samples. Constraints are created by drawing a set of polygons in the output video coordinate system. Polygons represent traversable regions, entry-exit regions and obstacles such as mailboxes. In-polygon testing on the trajectory samples in sidewalk coordinates is performed to decide acceptance or rejection.

Following unary constraint satisfaction, binary constraint satisfaction is performed by computing an independent set on a conflict graph as described in Section 4.4. The resulting set of trajectories is laid out as a Three-dimensional composition of numerous video billboards placed in front of a static background video billboard. As opposed to a Two-dimensional layer-based composition, a Three-dimensional layout enables easy visible surface determination and correct perspectivity.

4.1 Crowd Tubes for Constrained Video Billboard Composition

We introduce *crowd tubes*, our representation of video objects as a subvolume in the output video volume coupled with a trajectory of proxy shapes planted in a calibrated ground plane. Each shape serves as a proxy for a video object's ground plane occupancy for one frame. We assume that pedestrians are always in contact with a single ground plane and in our experiments, we represent ground occupancy with a circle for simplicity. When swept from the first frame to last, each circle trajectory forms a well-defined tube in the top-down rectified video volume (see Fig. 3), like a stack of coins.

Crowd tubes enable simple detection of collision and spatial separation constraint violations between a pair of crowd elements by performing distance thresholding on temporally corresponding circle proxies. Tighter-fitting proxy shapes, such as ellipses or bounded volume hierarchies, may also be employed for greater constraint violation detection accuracy. Crowd tubes are generated by translating and concatenating ground plane trajectories at transition points, thus forming a rigid path for Three-dimensional video billboards to follow. For each position in the path during animation, 1) the video billboard's object-centered coordinate system origin is updated to compensate for segmentation motion within the texture, and 2) the video billboard's origin is translated to its current Three-dimensional position (in the ground plane).

The artist authors a crowd by instancing crowd tubes that are anchored to specific points in time and space in the output video volume. A keyframe selected from a crowd tube's corresponding input clip V_i is displayed at the anchor point, thus providing for behavior control and a limited form of crowd choreography. Section 5.4 describes our interaction technique for anchoring and generating crowd tubes.

Extensions. In our experiments, we treated crowd tubes as rigid objects. However, crowd tubes can potentially be retimed and placed to achieve more variety in behavior. For example, stretching and squashing crowd tubes in the temporal direction correspond to slowing down and speeding up behavior playback. To limit noticeable frame replication or decimation, retiming should be constrained to stay within 10 percent of original playback rate. Behaviors may be recorded at high speed to decrease pedestrian speed without replicating or interpolating input frames. Furthermore, with additional analysis of segmented input video, crowd tubes could represent individual footsteps. This capability could support the accurate clustering of crowd tubes into supernodes to support the layout of entire lanes of pedestrian traffic or other aggregate crowd effects.

4.2 Unary Layout Constraints

The artist may specify polygons in the output view to represent a variety of layout constraints including traversable regions, entry-exit regions and obstacles. Polygons partition the output video volume into well-defined spaces that are used to accept or reject crowd tube samples as a preprocess for satisfying unary constraints before binary constraint satisfaction. Fig. 6 illustrates an example scene with two traversable polygons (drawn in green) corresponding to two sidewalks, and four entry-exit polygons (in red).

Given a crowd tube, which includes a concretely positioned rigid trajectory in sidewalk coordinates, in-polygon testing of each point in the trajectory yields a binary trajectory mask indicating in-out status per point. The mask serves two purposes: 1) to decide whether a crowd tube legally enters and exits the scene, traverses the scene in a designed traversable region and does not intersect with an obstacle that is part of the background clip V, and 2) to trim a crowd tube to only include trajectory points that lie inside a traversable region. The second purpose is important for efficient animation and rendering as extraneous concatenated segments of V_i would be needlessly animated and rendered.

A crowd tube must pass the following sequence of tests to be accepted before binary constraint satisfaction can occur. Testing is performed with trajectories and polygons mapped to sidewalk coordinates (see Section 5.3). First, a crowd tube's trajectory mask is computed via in-polygon testing with all entry-exit polygons. Note that an entry-exit polygon can serve as a location for both scene entry and exit. For example, a pedestrian may enter from the left side of the screen and leave on the right or vice versa for a pedestrian moving in the opposite direction. If the mask has two connected components, it satisfies the entry-exit constraint. Next, the crowd tube is trimmed to traversable regions by selecting a range within the trajectory from the first in-traversable-region point to the last traversable point and removing points outside the range. The crowd tube is trimmed a second time to include the range of points starting from the first entry-exit point to the last. Finally, a crowd tube is deemed to satisfy unary constraints if no trajectory point lies outside a traversable polygon.

4.3 Behavior and Density Control

In the previous section, we described how to accept or reject a crowd tube sample subject to unary layout constraints. In this section, we introduce an interaction technique for generating a seed set of crowd tubes of artist-controlled behavior and density.

Behavior. The goal is to enable an artist to specify when and where specific keyframes of V_i should appear in the output video. In the spirit of video textures, we assume that infinite playback of V_i is possible by computing transition points within V_i and walking the motion graph comprised of a node per frame and a directed edge per transition. Recent work [5] presented an in-studio approach to generating human video textures, or controllable animations made from joint video and motion capture of human motion. For the purpose of investigating the constrained layout challenges facing video-based crowd synthesis, we assume that each V_i has been preprocessed to identify transition frames and the motion graph is constrained to looping.

Behavior control is made possible by specifying a crowd tube's anchor point in the output video volume. An anchor point is a four-tuple (x, y, t_{out}, t_{in}) specifying that frame t_{in} of V_i should appear in the output video volume at (x, y, t_{out}). Given an anchor point, a crowd tube is generated by concatenating frames both forward and backward in time according to the motion graph associated with its clip V_i. In principle, multiple anchor points may be established to animate a crowd element using multiple keyframes. In practice, forwards-backwards concatenation is performed from a single anchor point using a random walk on the motion graph (or a simple loop). This simple keyframe-based approach to behavior control can enable a variety of tightly choreographed crowd outputs.

Density. How should an artist control the density of a video-based crowd? In model-based crowd synthesis, the crowd artist typically establishes an emitter location and flow rate in the style of particle systems. Agent-based simulation ensures that all emitted crowd elements satisfy unary and binary layout constraints by virtue of the agent's dynamical model. In our video-based approach, the number and proximity of constraint satisfying crowd tubes are governed by the process of generating crowd tube candidates before constraint satisfaction, which we define as the *seed set*.

We loosely provide for density control by treating the seed set as an upper bound for the population and density of the crowd result. It is possible for all crowd tubes in a seed set to satisfy all constraints defined in this paper, but this is unlikely to occur in space-time regions of high density. Therefore, the video-based crowd artist can approximate a desired density by generating a seed set whose density is proportional to the expected constraint-satisfying set. A high concentration of anchor points (and consequent crowd tubes) in a space-time region of the output will loosely define a desired density.

To support the otherwise tedious task of establishing an anchor point for each sample in the seed set, we created a

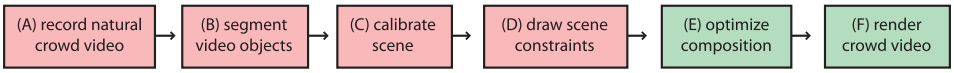


Fig. 4. Pipeline for video-based crowd synthesis. Red boxes indicate steps requiring user interaction and green boxes are automatic.

"spray" interaction technique to distribute a set of crowd tubes with user-defined center and spread in the output video volume. The artist selects one of four bins of crowd tubes (1, 5, 50, or 100) to spray at a time. A continuous slider was excluded from the interface for increased choreography speed. A specific V_i may be selected or randomly chosen for spraying. The keyframe t_{in} corresponding to V_i and spread in terms of standard deviation from t_{in} in frames are specified in the interface for normally distributing crowd tubes in the output video volume. A single click sprays the normally distributed crowd tubes with a spatial variance equal to the squared density (e.g., 25 for a spray of five tubes). Section 5.4 presents our interface for spraying.

4.4 Constraint Optimization Problem

After generating a seed set of crowd tubes, we must solve a constraint optimization problem: Given a seed set of crowd tubes, select a constraint-satisfying subset of tubes of maximal cardinality. The cardinality maximization objective is important for controlling crowd population and density; as cardinality increases, the gap between the expected density and actual density decreases. As the artist increases the density of crowd tubes to occupy all free space in the output volume, the constraint optimization problem becomes a packing problem.

4.5 Conflict Graph and Independent Set

To solve this general constraint optimization problem, we observe that it can be reduced to a more specific computational task on a graph: the maximal independent set (MIS) problem. MIS is defined as the largest subset of nodes in a graph such that no pair share an edge. MIS is equivalent to computing the maximal clique in the graph's complement and is known to be NP-hard [29].

Given a set of crowd tubes, we construct a conflict graph G with a node per crowd tube and an edge per binary constraint violation. The seed set is preprocessed to instantiate a graph node for each crowd tube that satisfies unary constraints, as previously described in Section 4.2. An edge is inserted between each pair of nodes p and q if at least one of the following binary constraint violations occurs:

- A collision is detected at any point in time between p and q, which occurs if the L2 distance between p and q in sidewalk coordinates falls below a collision detection threshold t_c.
- If p and q are generated using the same input clip V_i (defining p and q as *clones*), the L2 distance between p and q in sidewalk coordinates falls below a spatial proximity threshold t_p.

Following conflict graph construction, we compute an approximate maximal independent set I using Luby's algorithm [29]. The solution set of crowd tubes I is then used to animate and render a video-based crowd. Animation and rendering details are presented in Section 5.5.

5 System

The crowd tube-based representation, constraint authoring interaction techniques, constraint satisfaction algorithm, and rendering procedure were used to construct an interactive system for synthesizing video-based crowds. Fig. 4 illustrates the linear pipeline. This section describes each of the steps in detail. The following section describes four experiments using our system and the paper concludes with a discussion of our findings.

5.1 Capture

To address the primary research question of how to layout segmented video objects while satisfying unary and binary constraints imposed by the Three-dimensional scene, we designed a video capture process to meet three objectives: 1) the crowd artist should be free to casually record video without having to direct actors to move in specific ways, 2) the crowd artist should not be required to insert specially constructed calibration objects into the scene, and 3) the artist should be able to pan, tilt and zoom into the scene to capture an observed behavior with maximal resolution. The camera's motion is restricted to panning, tilting, and zooming to support simple homography-based calibration, which is described in Section 5.3. The scene is also restricted to contain a single primary ground plane where synthetic crowds are to be added. In principle, our technique is extendable to multiple planes or, with a more complicated calibration process, to enable crowd tube trajectories to lie on the surface of arbitrary geometry. Crowd synthesis on rolling hills, for example, would require hill geometry and potentially more video clips to account for the greater variation in dynamics as individuals climb up or speed down hills.

In our experiments, we captured four scenes using a tripod-mounted high definition video camcorder (specifically, the Sony VIXIA HFS100 model). Three of the scenes were recorded in an undirected manner for the purpose of populating the scene with natural behaviors. In the fourth scene, we directed an actress to validate the choreography process in a planned manner. Across all scenes, the capture process consisted of choosing a suitable vantage point, selectively zooming and locking in on desirable behaviors for synthesis, and choosing a background video. In principle, video may be captured while continuously panning, tilting, and zooming. In practice, we chose to fix the camera's pose after adjusting the field of view separately for each clip V_i and for the background plate V. In future work, the camera's motion may be stabilized using automatically recovered homographies.

5.2 Segmentation

After recording the scene and its crowd elements, which include individuals, groups of individuals, and inanimate objects, each crowd element's source clip V_i must be segmented for ground plane trajectory estimation and video billboard matting. An accurate, temporally coherent segmentation result is necessary to create the illusion of a plausible synthetic crowd. Furthermore, a rapid, interaction-limited segmentation process is required to generate a variety of crowd tube clip sources V_i in a practical amount

TABLE 1
Segmentation Interaction Time: This Table Reports Time Spent on Interactive Video Stabilization and Segmentation for Several Input Clips in the Embarcadero Scene

sequence	image	frames	stabilization	segmentation
bluesuitman	Fig 8b	233	0 min.	47 min.
womanleft	Fig 8d	240	8	245
groupof4	Fig 8e	340	7	215
roundtable	Fig 8f	77	5	24

Note that segmentation time is not proportional to the clip duration, but rather a function of the image content. Challenges to segmentation include foreground-background color overlap and large shape changes.

of time. There is a tradeoff between artifacts associated with segmenting many clips of short duration and segmenting a lesser number of clips with longer duration. Shorter clips increase the frequency of transitions required to animate a crowd element moving from an entry to an exit (or staying in place for inanimate objects). On the other hand, longer clips display transitions less frequently in the output but at the cost of lowering the quantity of behavior sources V_i, thus potentially damaging the perception of crowd variety.

To achieve a greater level of accuracy at the expense of interaction time, we segmented the clips using a commercially available interactive video object segmentation system. Specifically, we used the "Rotobrush" tool in Adobe After Effects CS5. While the technique is unpublished, a number of interactive segmentation papers have been recently published by the Adobe company [30], [31]. Figs. 7, 8, 9, and 10 illustrate example background views V and segmented behaviors from each V_i as time-lapse visualizations. Table 1 reports the amount of interaction time required to segment several example clips in the *Embarcadero* scene, displayed in Fig. 8.

5.3 Calibration

Calibration is necessary to animate video billboards on a Three-dimensional ground plane in the scene. To summarize, calibration enables the following three step process for constructing and animating crowd tubes. First, ground tracks extracted from each segmented clip V_i are mapped to sidewalk coordinates. Second, a crowd tube and its entry-to-exit trajectory are generated by concatenating segments of the ground track in the sidewalk frame. Finally, crowd tube sidewalk trajectories are mapped to the output view and raycasting produces a Three-dimensional entry-to-exit trajectory planted in the ground plane. The three coordinate frames are exploited during the crowd tube construction process, illustrated in Fig. 5.

Our interactive process consists of two main steps: 1) estimation of ground plane parameters A, B, C, D corresponding to the $Ax + By + Cz + D = 0$ representation, and 2) estimation of a series of homographies: output-sidewalk homography $\mathbf{H_{os}}$ (1 per scene) and input-output homography $\mathbf{H_{io}}$ (1 per clip V_i). During animation, a homogeneous Two-dimensional point $\mathbf{x_i}$ in the zoomed-in input frame V_i is mapped to a point $\mathbf{x_o} = \mathbf{H_{io}}\mathbf{x_i}$ in the output view V. Given a virtual camera placed at the Three-dimensional origin with principal point $\mathbf{x_c}$ (set to the center of V) and focal length f (set arbitrarily), a Three-dimensional ray is constructed with a direction toward output point $\mathbf{x_o}$ and cast through the calibrated ground plane to retrieve the billboard's Three-dimensional position $\mathbf{X}$ per crowd tube and per output frame.

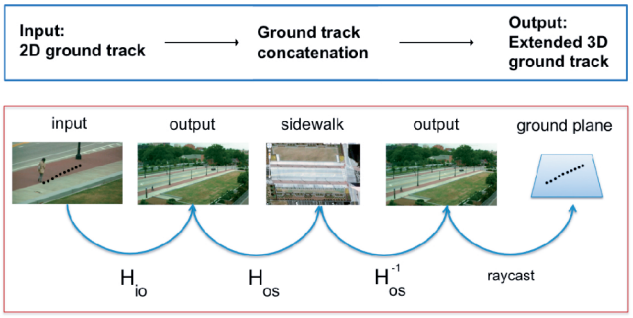


Fig. 5. Trajectory concatenation to generate a crowd tube. Given an input video clip with a segmented behavior (woman on left wearing beige in this example), the Two-dimensional ground track is extracted automatically and mapped to sidewalk coordinates using calibrated homographies. Sidewalk trajectories are concatenated to extend the track from scene entry to exit. Following concatenation, the extended track is mapped to Three-dimensional ground plane coordinates by raycasting the trajectory's output coordinates through a calibrated ground plane.

Ground plane. We simplify the interactive ground plane calibration process by making a reasonable assumption about the scene: at least four objects of approximate fixed height, such as lamp posts or adult humans, are standing orthogonal to the ground plane. The object r that is furthest away is chosen as a *reference* whose base point is fixed at $z = f$ and the remaining objects' Three-dimensional base points are estimated relative to r. Based upon this assumption, we solve for $\mathbf{x} = [A, B, C, D]^T$ parameters in a least-squares manner: Find $\mathbf{x}$ that minimizes $\|A\mathbf{x}\|$ where

$$A = \begin{bmatrix} s^1 x_b^1 & s^1 y_b^1 & s^1 f & 1 \\ s^i x_b^i & s^i y_b^i & s^i f & 1 \\ \vdots & \vdots & \vdots & 1 \\ s^n x_b^n & s^n y_b^n & s^n f & 1 \end{bmatrix},$$

and where $s^i = \|x_b^r - x_t^r\| / \|x_b^i - x_t^i\|$ is the ratio of the reference object r's height to nonreference object i's height in the output frame V. User interaction in the form of at least eight mouse clicks on V provides feature coordinates corresponding to the bottom point $[x_b^i\ y_b^i]^T$ and top point $[x_t^i\ y_t^i]^T$ for four or more objects in principal point centered coordinates $\mathbf{x} - \mathbf{x_c}$. At least three or more nonreference objects' Three-dimensional base points are determined relative to r using the fixed height and orthogonality assumptions. The objects should be distributed to span the area where synthetic crowds are desired, as opposed to concentrated in one small region of the area, to avoid errors caused by overfitting.

Homographies. The user supplies four point correspondences between each V_i and V to estimate $\mathbf{H_{io}}$ using direct linear transformation [32]. An additional set of four point correspondences are manually specified to estimate $\mathbf{H_{os}}$ for mapping output view V to sidewalk frame V_s. As shown in Fig. 5, we obtain top-down views V_s of each scene using publicly available aerial images. There are several techniques for estimating metric-rectified homographies [32] if an indoor scene or outdoor scene lacking aerial imagery is desired for crowd synthesis. We chose a simpler approach

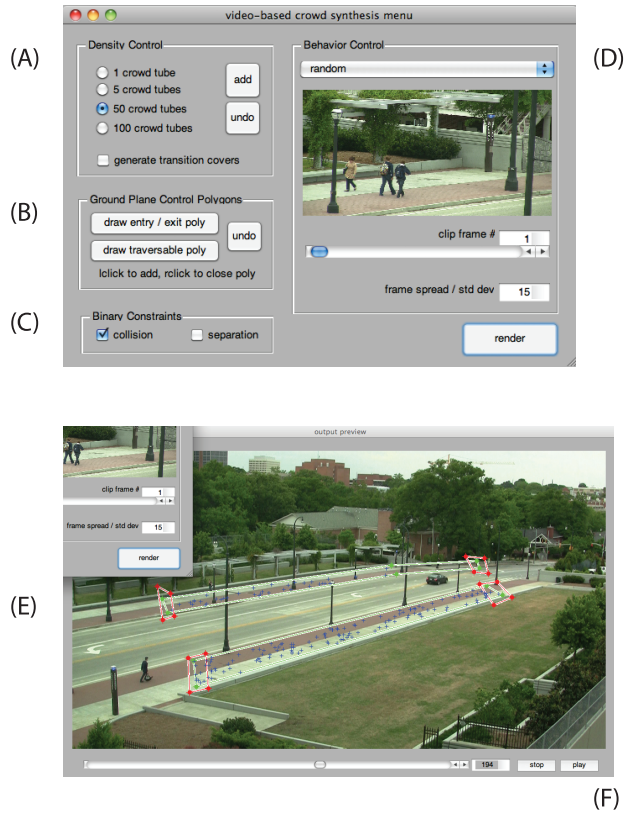


Fig. 6. Crowd authoring interface. A) Four density levels are available for distributing crowd tubes in the output as a batch. B) Polygons represent entry-exit regions and traversable regions for constrained crowd tube placement. C) Collision avoidance and spatial separation of clones are two binary constraints which can be activated before synthesizing a video-based crowd. D) A user may choose from a set of segmented input clips V_i for instantiating crowd tubes. E) Unary constraint polygons and crowd tube positions per frame may be previewed in the output window. F) An output view's time slider allows for crowd tube placement at specific points in time and crowd previsualization, where each crowd tube's position is marked with a blue cross.

as most metric rectification techniques require additional information, such as five or more orthogonal line pairs planted in the plane.

5.4 Crowd Authoring

After capturing, segmenting, and calibrating a scene, the artist designs a video-based crowd by annotating the output view V with a set of polygons followed by crowd tube instantiation. We constructed a crowd authoring interface (see Fig. 6) to enable annotation and to give the user a basic previsualization of their crowd. A typical crowd authoring session involves the following procedure. First, the user draws polygons representing traversable regions where crowds may lie. Likewise, entry and exit zones are demarcated with polygons. The green polygon in Fig. 6 represents an example traversable region and the red boxes indicate entries and exits, whose meaning depends on a crowd tube's direction of travel. Polygons serve as unary constraints which act on rigid trajectory samples, as previously described in Section 4.2. Second, a crowd is constructed by clicking in V to *spray* crowd tubes into the scene, thus enabling a limited form of behavior and density control over a crowd composition. Finally, the system solves the constraint satisfaction problem (see Section 4.4) and animation and rendering follows.

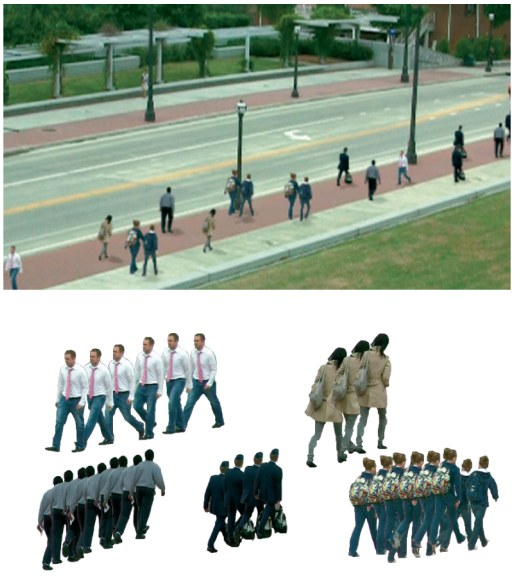


Fig. 7. Bridge result and data set.

Fig. 8. Embarcadero result and data set. Interactive segmentation timings for clips B, D, E, and F are provided in Table 1.

5.5 Animation and Rendering

To support the computationally expensive process of animating and texture mapping high definition video billboards against a video background plate, we built our system on top of an industry-standard video composition system. Following crowd authorship and constraint optimization, our system generates a script which is then processed by Adobe After Effects CS5 for automatic animation and rendering of output video. Before script execution, the user must set up an environment in the commercial application with a list of video assets and corresponding matte frames associated with each input source V_i. Script execution and rendering is a batch, offline process that takes between 1 and 4 hours in our experiments. The artist may exploit several features provided by the commercial system to improve the look of the crowd. For example, we produced virtual shadows

Fig. 9. Sailboat result and data set.

for the *Bridge* scene by rotating, blurring, scaling, and alpha attenuating segmentation clips.

6 Data Set and Results

Using our system, we produced a series of video-based crowds to demonstrate the capabilities of our crowd tube representation, constraint satisfaction algorithm, and synthesis pipeline on four outdoor scenes. Our primary experimental objective was to answer the following research question: Given a video of a sparsely crowded scene, how can we synthesize a plausible video of the same scene populated with crowds of artist-controlled density and behavior? We generated a range of video outputs that showcase crowds before and after enforcing collision and variety constraints. Results are provided for scenes with virtual shadows (*Bridge*), with real, segmented shadows (*Embarcadero*) and without shadows (*Sailboats* and *Park*). In this section, we report detailed observations for each sequence.

Bridge. The first sequence in the results video (see Fig. 7) depicts a result from the *Bridge* scene and corresponding data set, which is comprised of five video clips V_i, a crowd-free background clip V and calibrated ground plane and homographies. Results demonstrate compositional quality including perspective correctness, segmentation accuracy, traversability satisfaction, lighting and shading, and crowd variety. Fig. 7 reveals a crowd of two women in beige (two moving pedestrians segmented together), two men in a bright shirt, two pairs of friends walking together, three men carrying a backpack, and three men in gray sweatshirts. The difference in height between the man in the beige sweatshirt on the left and the same man far on the right reveal a correct perspective effect. No pieces of the body appear to be missing due to segmentation artifacts. Pedestrians remain on sidewalks as designated by control polygons that apply unary constraints on crowd tubes at authoring time (see Fig. 6b). Binary crowd tube constraints are also noticeably enforced; none of the crowd elements appear to collide, or occupy the same space in the ground plane. Clones are spatially separated. These issues demonstrate that basic compositional objectives are being met.

They also show how our video-based crowd synthesis system may be used to control a large number of synthesis variables with a tight level of control. Thus, the system lends itself to extensive perceptual studies in the spirit of recent perceptual studies of factors that affect crowd realism [28], [33]. In contrast with those works, which employ nonphotorealistic model-based avatars, our system presents the opportunity to control and investigate more subtle properties of photo-realism and video-realism for the first time.

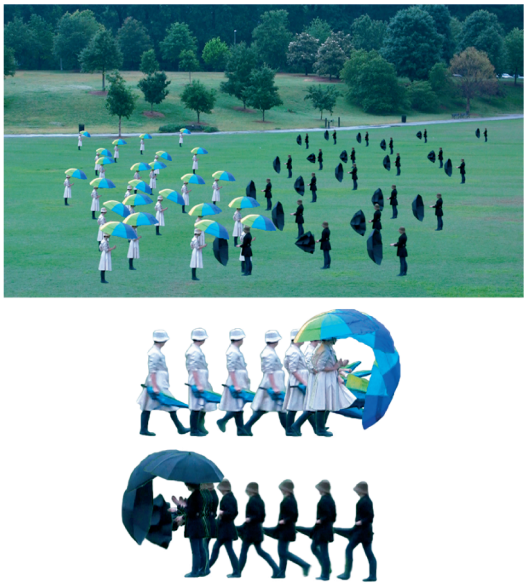


Fig. 10. Park result and data set.

Embarcadero. We designed a second data set, titled *Embarcadero*, to observe the effect of increasing the quantity, duration, and variety of crowd elements. In contrast with *Bridge*, a major pathway for students crossing campus, *Embarcadero* is a popular tourist destination in San Francisco, thus presenting more opportunities to capture casual behaviors of visitors. Fig. 7 presents a still from a *Bridge* result and corresponding data set, which consists of seven clips spanning a range of dynamics. Three of the input clips (see Figs. 8a, 8b, and 8d) contain linear walking trajectories. Fig. 8g shows a tourist holding a camcorder while standing still, followed by walking away after recording. The remaining clips include static crowd elements including: Fig. 8e shows a group of four visitors who have stopped so that one man can place his sweater in another's backpack, Fig. 8f shows two women having a discussion at a round-table, and Fig. 8c shows a woman texting while seated on the ground. Table 2 provides crowd tube counts and crowd authorship, animation, and rendering timings.

Sailboats. A third scene titled *Sailboats* was selected to investigate the issue of crowd output quality versus authorship time. While *Embarcadero* depicts a wide variety of complex behaviors at the cost of approximately one full day of interactive segmentation work and rendering time, *Sailboats* includes only two light-colored sailboats moving in opposite directions against a blue background (see Fig. 9) placed at a distance. A crowd of two sailboats was constructed from capture to render time in less than 3 hours of work. Segmentation required little to no interaction, likely due to little shape change and foreground-background overlap.

Park. Finally, we captured and processed a fourth scene (see *Park* illustrated in Fig. 10) as an example of planned crowd choreography. Inspired by typical crowd effects involving battling armies, we aimed to produce a "battle" of

TABLE 2
Crowd Tube Sample Counts Before and After Constraint Satisfaction: The Authoring Interface Was Used to Quickly Generate a Random Crowd in 40 Seconds by Spraying Batches of 100 Random Crowd Tubes per Mouse Click

sequence	sample count	sample count in MIS
beigepants	97	23
cameraman	77	8
bluesuitman	101	25
groupof4	8	0
textingwoman	1	0
womanleft	74	5
roundtable	3	0
total	361	61

Of 361 unary constraint-satisfying crowd tubes, Luby's algorithm computed an approximate MIS of cardinality 61 in 50.44 seconds. Animation and rendering took approximately 1 hour 40 minutes for a 10 seconds result.

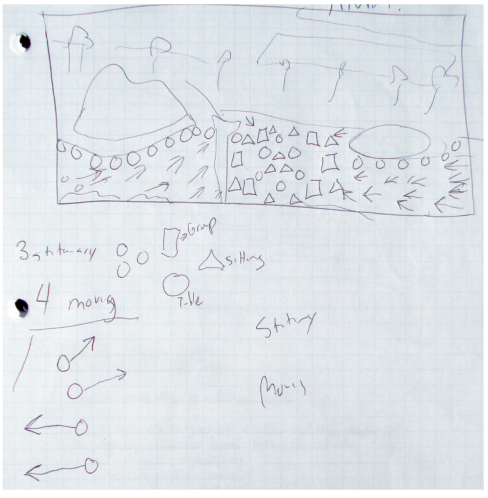


Fig. 11. Crowd plan for User 1. The user intended to have a large collection of static objects in the center while hordes of pedestrians marched from both sides toward the center while slipping through the obstacles upon arrival.

umbrella-wielding individuals who approach each other as a marching army and stop short of clashing to open an umbrella. We varied the shape of the line of both moving fronts at umbrella-opening time to demonstrate our system's behavior control capabilities. Umbrella-opening keyframes were instantiated with a standard deviation of 15 frames (using interface controls displayed in Figs. 6d and 6f) into the output video volume in a designed fashion. Fig. 10 displays the result for the *Park* scene. Video results demonstrate our system's ability to choreograph a behavior-controlled crowd of this nature.

7 Qualitative User Evaluation

In addition to generating a range of crowds across four scenes to demonstrate the capabilities of our prototype system, we designed and conducted a crowd authoring exercise to assess the system's usability, bottlenecks and favored features. The primary goal was to qualitatively investigate how effective the system is in allowing someone to realize their "vision" of a video-based crowd animation with little to no prior experience with the system. Three users who had previous experience with advanced video editing software were asked to plan, author and evaluate a

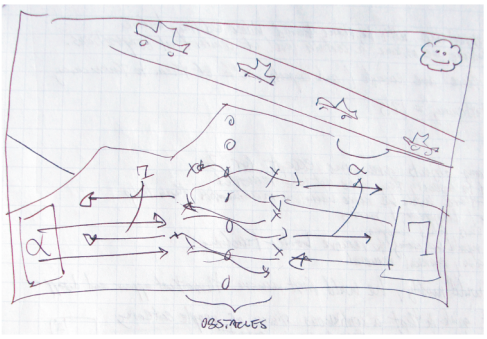


Fig. 12. Crowd plan for User 2. Similar to User 1's goal, User 2 planned a crowd which would exhibit pedestrians marching through a line of static obstacles.

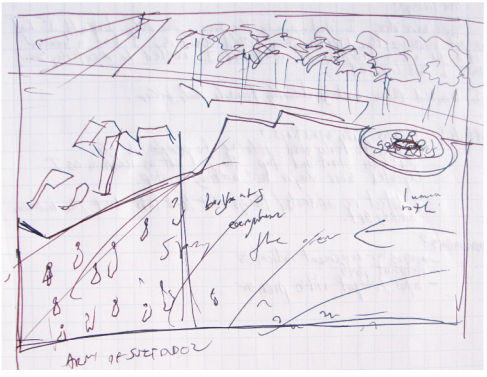


Fig. 13. Crowd plan for User 3. The user's intent was "to embrace the surreal and create a large group of suit-wearing men walking to a woman moving in an opposing direction to see if the woman disappeared among them."

Fig. 14. Select output frame for User 1. The output indicates that the general goal described in Fig. 11's caption was met; a large section of immobile behaviors are exhibited in the center while groups march in from the sides.

video-based crowd using the *Embarcadero* data set (see Figs. 11, 12, 13, 14, 15, and 16).

Design rationale. By constraining the users to create crowds within 30 minutes and with a limited palette of seven behaviors for a single scene, we hoped to discover positive and negative trends across all users' experiences. Specifically, we wanted to uncover common planning and design strategies for coping with the same limited set of input examples.

7.1 User Trends

In this section, we highlight several common problems and desired features among all users who evaluated the system.

Fig. 15. Select output frame from User 2. Constraint satisfaction removed all but three static obstacles positioned near the center of the frame. The result did not meet the expectations in the corresponding sketch exhibited in Fig. 12.

Fig. 16. Select output frame for User 3. The result met User 3's expectations.

We extracted the problems and feature ideas from think aloud comments and questions asked during the authorship and evaluation steps. Some of the trends include specific observations for new features and improvements to the process and interface for authoring a video-based crowd animation.

- *Iterative editing and constraint satisfaction.* Early on in the authoring process, all three users expressed a desire to preview their crowd output after constraint satisfaction had occurred. The prototype system does not currently support iterative adding, solving, and previewing the crowd in progress and this proved to be the most important limitation for the users.
- *Behavior inspection and organization.* All three users planned their crowd by first reviewing the available behaviors and sorting them by degree of motion (static versus moving along ground). Also, all three users asked to repreview each behavioral clip to understand how the element will move and made note of the motion on a sheet of paper. They also all suggested adding arrow icons below each clip preview (see Fig. 6d, at top) to view a summary of the clip's motion without having to watch it.
- *Crowd tube layout visualization.* Two of the users suggested adding "footprint" icons indicating the full trajectory of a crowd tube as the mouse is moved inside the traversable region.
- *Easy to learn.* All three users, who had never used the system before, expressed satisfaction with their ability to synthesize their own crowd video starting from a hand-drawn sketch within 30 minutes.

Based on our findings, we believe there are two crucial areas of improvement for future work on the crowd authoring processing and interface: 1) behavior categorization and visualization features would partially address the primary issues raised concerning crowd planning and behavior selection and 2) having the human in the loop with iterative constraint satisfaction would ease concerns regarding the difference between planned and constraint-satisfied crowds.

8 Limitations

In addition to the specific crowd authoring interface issues uncovered in the previous section, there are several limitations to general video-based crowd synthesis.

First and foremost, the variety of crowd outputs is limited to behaviors observed at capture time. This is a limitation facing image-based rendering techniques in general. While increasing the size of the data set would alleviate the problem, it comes at the expense of interactive segmentation and calibration time.

Data set limitations may be partially addressed by modulating the color of pedestrian clothes with additional segmentation work. Researchers have previously demonstrated the effectiveness of increasing crowd variety using color modulation [33]. We demonstrate initial color modulation results in the final clip of our result video. A second, complementary approach to color modulation is to increase the length of captured behaviors by trading resolution for time and zooming out. Longer sequences provide more transition opportunities, leading to an exponential increase in behavior branching and more varied, nonlinear crowd tubes. Furthermore, crowd tubes may be squashed and stretched in the temporal direction within clip retiming thresholds to further exploit the limited collection of behaviors. We empirically determined retiming thresholds and set them at 10 percent of captured playback rate. Finally, collision and variety constraints may be switched off to increase density, which may appear perceptually plausible depending on the specific crowd layout and group size. Online video results illustrate this effect.

Second, transitions artifacts appear in the form of popping artifacts as our current system uses basic jump cuts between frames out of sequence. Transition synthesis techniques [5] address the problem but they currently require an elaborate motion capture setup against a blue screen backdrop. An alternative technique for animating people textures [34] may be applied for sprites with the limitation that they remain in place but without requiring motion capture. Moving gradients [35], a nonrigid frame interpolation technique, could potentially address the limitation.

Third, crowds are animated atop a single ground plane in our prototype system, thus eliminating a large category of interesting scenes. This may be alleviated by calibrating multiple planes within a scene. Additional forms of proxy geometry, such as a landscape composed of Bézier patches, may also be employed with the added complication of layering sprites when occluded (e.g., when a group walks over a hill). Note that hills would be limited in their degree of incline as large hills would present varying visibility of moving pedestrians. A second, related limitation to ground plane geometry is the fixed viewpoint of the camera. A fixed view of proxy geometry for the crowd plus a moving

collection of sprites may provide for zooming with parallax motion of pedestrians, as demonstrated in the results video. However, a moving camera would require a multiview capture as recently demonstrated [36].

Virtual agents versus video billboards. Despite these limitations, the proposed technique is well suited for repopulating scenes with zoomed in detail over a large volume of crowd elements. Close-up shots can reveal the synthetic nature of virtual agents shown in previous work (e.g., [1], [28]). The final video clip in our result video gives an example of output that would take considerably more time to produce with crowd agents than with our example-based method.

9 Conclusion and Future Work

This paper described how to synthesize video-based crowds by animating a collection of matted video billboards. We presented a constraint satisfaction problem and solution for placing video billboards in a Three-dimensional scene subject to constraints on motion and appearance. We designed a system and algorithm for composing a constraint-satisfying set of *crowd tubes*, our representation of segmented videos coupled with a trajectory of ground occupancy proxy shapes (e.g., circles) planted in a ground plane. The presented results portray three natural, unchoreographed scenes and one planned and directed crowd. In addition to providing a range of crowd outputs from four scenes, a qualitative user evaluation of the prototype system uncovered issues, and directions for future work on the crowd authoring process and interface.

There are several exciting avenues for future work in video-based crowd synthesis:

- Large databases of high-quality blue screen-matted video clips are becoming commercially available, presenting opportunities for large scale video-based crowd synthesis. This presents several new challenges including how to automatically transfer matted clips across scenes while accounting for variations in lighting and appearance. It also presents computational demands on the present constraint satisfaction approach, which is NP-hard.
- A design gallery-based approach to authoring crowds could enable a higher level approach to controlling captured behaviors at the individual and macro scales. As discovered during the qualitative user evaluation (Section 7), user-controlled behavior grouping and categorization would potentially ease frustrations with the crowd authoring process.
- Crowd synthesis packages, such as Massive [6], could incorporate the crowd tube representation and optimization technique for rapid reuse of previously animated and rendered behaviors.

Acknowledgments

The authors would like to thank Dr. Kiran Bhat and Professors Karen Liu, Irfan Essa, and Sing Bing Kang for many useful discussions. This work was supported in part by ONR HUNT MURI. Matthew Flagg did this work as a PhD student, and he is now CTO of PlayVision Labs, Inc.

References

[1] A. Treuille, S. Cooper, and Z. Popović, "Continuum Crowds," *Proc. ACM SIGGRAPH,* pp. 1160-1168, 2006.

[2] J. Ondřej, J. Pettré, A.-H. Olivier, and S. Donikian, "A Synthetic-Vision Based Steering Approach for Crowd Simulation," *ACM Trans. Graphics,* vol. 29, pp. 123:1-123:9, July 2010.

[3] E. Ju, M.G. Choi, M. Park, J. Lee, K.H. Lee, and S. Takahashi, "Morphable Crowds," *ACM Trans. Graphics,* vol. 29, 2010.

[4] A. Schödl and I.A. Essa, "Controlled Animation of Video Sprites," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA),* pp. 121-127, 2002.

[5] M. Flagg, A. Nakazawa, Q. Zhang, S.B. Kang, Y.K. Ryu, I. Essa, and J.M. Rehg, "Human Video Textures," *Proc. Symp. Interactive Three-dimensional Graphics and Games (IThree-dimensional '09).* pp. 199-206, 2009.

[6] M. Software, "Massive, Crowd Animation Software for Visual Effects," http://www.massivesoftware.com, 2003.

[7] B. Hiebert, J. Dave, T.-Y. Kim, I. Neulander, H. Rijpkema, and W. Telford, "The Chronicles of Narnia: The Lion, the Crowds and Rhythm and Hues," *Proc. ACM SIGGRAPH,* p. 1, 2006.

[8] J. Funge, X. Tu, and D. Terzopoulos, "Cognitive Modeling: Knowledge, Reasoning and Planning for Intelligent Characters," *Proc. ACM SIGGRAPH,* pp. 29-38, 1999.

[9] W. Shao and D. Terzopoulos, "Autonomous Pedestrians," *Graphical Models,* vol. 69, nos. 5/6, pp. 246-274, 2007.

[10] O.B. Bayazit, J. Lien, and N.M. Amato, "Better Group Behaviors in Complex Environments Using Global Roadmaps," *Proc. Eighth Int'l Conf. Artificial Life,* pp. 362-370, 2002.

[11] D. Helbing, L. Buzna, A. Johansson, and T. Werner, "Self-Organized Pedestrian Crowd Dynamics: Experiments, Simulations, and Design Solutions," *Transportation Science,* vol. 39, no. 1, pp. 1-24, 2005.

[12] D. Helbing, P. Molnar, I.J. Farkas, and K. Bolay, "Self-Organizing Pedestrian Movement," *Environment and Planning B: Planning and Design,* vol. 28, pp. 361-383, 2001.

[13] C. Loscos, D. Marchal, and A. Meyer, "Intuitive Crowd Behaviour in Dense Urban Environments Using Local Laws," *Proc. Theory and Practice of Computer Graphics Conf. (TPCG '03),* pp. 122-129, 2003.

[14] R.A. Metoyer and J.K. Hodgins, "Reactive Pedestrian Path Following from Examples," *Proc. 16th Int'l Conf. Computer Animation and Social Agents (CASA '03),* pp. 149-156, 2003.

[15] M. Sung, "Scalable, Controllable, Efficient and Convincing Crowd Simulation," PhD dissertation, Madison, WI, USA, supervisor-Gleicher, Michael L, 2005.

[16] C.W. Reynolds, "Flocks, Herds and Schools: A Distributed Behavioral Model," *Proc. ACM SIGGRAPH,* pp. 25-34, 1987.

[17] M. Sung, L. Kovar, and M. Gleicher, "Fast and Accurate Goal-Directed Motion Synthesis for Crowds," *Proc. ACM SIGGRAPH/ Eurographics Symp. Computer Animation (SCA '05),* pp. 291-300, 2005.

[18] L. Kovar, M. Gleicher, and F. Pighin, "Motion Graphs," *ACM Trans. Graphics,* vol. 21, no. 3, pp. 473-482, 2002.

[19] J. Lee, J. Chai, P.S.A. Reitsma, J.K. Hodgins, and N.S. Pollard, "Interactive Control of Avatars Animated with Human Motion Data," *ACM Trans. Graphics,* vol. 21, pp. 491-500, July 2002.

[20] B. Yersin, J. Maïm, J. Pettré, and D. Thalmann, "Crowd Patches: Populating Large-Scale Virtual Environments for Real-Time Applications," *Proc. Symp. Interactive Three-dimensional Graphics and Games (IThree-dimensional '09).* pp. 207-214, 2009.

[21] K.H. Lee, M.G. Choi, and J. Lee, "Motion Patches: Building Blocks for Virtual Environments Annotated with Motion Data," *ACM Trans. Graphics,* vol. 25, no. 3, pp. 898-906, 2006.

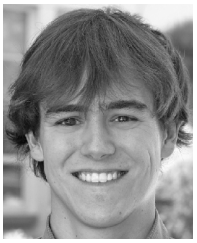
[22] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by Example," *Computer Graphics Forum,* vol. 26, no. 3, pp. 655-664, 2007.

[23] K.H. Lee, M.G. Choi, Q. Hong, and J. Lee, "Group Behavior from Video: A Data-Driven Approach to Crowd Simulation," *Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA '07),* pp. 109-118, 2007.

[24] N. Courty and T. Corpetti, "Crowd Motion Capture," *Computer Animation Virtual Worlds,* vol. 18, pp. 361-370, Sept. 2007.

[25] F. Tecchia, C. Loscos, and Y. Chrysanthou, "Image-Based Crowd Rendering," *IEEE Computer Graphics Application,* vol. 22, no. 2, pp. 36-43, Mar. 2002.

[26] L. Kavan, S. Dobbyn, S. Collins, J. Zara, and C. O'Sullivan, "Polypostors: Second Polygonal Impostors for third Crowds," *Proc. ACM SIGGRAPH,* pp. 149-155, Feb. 2008.
[27] J.-F. Lalonde, D. Hoiem, A.A. Efros, C. Rother, J. Winn, and A. Criminisi, "Photo Clip Art," *ACM Trans. Graphics,* vol. 26, no. 3, Aug. 2007.
[28] R. McDonnell, M. Larkin, S. Dobbyn, S. Collins, and C. O'Sullivan, "Clone Attack! Perception of Crowd Variety," *Proc. ACM SIGGRAPH,* pp. 1-8, 2008.
[29] M. Luby, "A Simple Parallel Algorithm for the Maximal Independent Set Problem," *Proc. 17th Ann. ACM Symp. Theory of Computing (STOC '85),* pp. 1-10, 1985.
[30] X. Bai, J. Wang, D. Simmons, and G. Sapiro, "Video Snapcut: Robust Video Object Cutout Using Localized Classifiers," *Proc. ACM SIGGRAPH,* pp. 1-11, 2009.
[31] B.L. Price, B.S. Morse, and S. Cohen, "Livecut: Learning-Based Interactive Video Segmentation by Evaluation of Multiple Propagated Cues," *Proc. IEEE 12th Int'l Conf. Computer Vision (ICCV),* 2009.
[32] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision,* second ed. Cambridge Univ. Press, 2004.
[33] R. McDonnell, M. Larkin, B. Hernández, I. Rudomin, and C. O'Sullivan, "Eye-Catching Crowds: Saliency Based Selective Variation," *Proc. ACM SIGGRAPH,* pp. 1-10, 2009.
[34] B. Celly and V. Zordan, "Animated People Textures," *Proc. 17th Int'l Conf. Computer Animation and Social Agents,* 2004.
[35] D. Mahajan, F.-C. Huang, W. Matusik, R. Ramamoorthi, and P. Belhumeur, "Moving Gradients: A Path-Based Method for Plausible Image Interpolation," *Proc. ACM SIGGRAPH,* pp. 42:1-42:11, 2009.
[36] F. Xu, Y. Liu, C. Stoll, J. Tompkin, G. Bharaj, Q. Dai, H.-P. Seidel, J. Kautz, and C. Theobalt, "Video-Based Characters: Creating New Human Performances From a Multi-View Video Database," *ACM Trans. Graphics,* vol. 30, pp. 32:1-32:10, Aug. 2011.



Matthew Flagg received the PhD degree in computer science from the Georgia Institute of Technology, Atlanta. He is the Chief Technology Officer and cofounder of PlayVision Labs, a startup focused on computer vision for entertainment. His research interests include computer vision, graphics, and human-computer interaction. He is a member of the IEEE.

James M. Rehg received the PhD degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA. He is a Professor in the College of Computing at the Georgia Institute of Technology. He is Associate Director of Research in the Center for Robotics and Intelligent Machines and co-Director of the Computational Perception Lab. His research interests include computer vision, robotics, machine learning, and computer graphics. He is a member of the IEEE.