Crumby Computer Assistance for Annotation Tasks

Anonymous for Review

ABSTRACT

We present our iterative design of an interface to assist users annotating moving objects. We also provide details of a user study of twenty participants using our interface to annotate video of ice-hockey players. Although we found computer assistance reduced the performance of users, we think our interface design is novel and that our results compel researchers to provide users with control over computer assistance during annotation. These contributions will be of interest to researchers in a broad range of annotation domains, such as human behaviour research, surveillance and augmented reality.

Author Keywords

Annotation, Object Positioning, Tracking, Interaction, Computer Vision, Shared Control

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION

Researchers think of annotation as details tagged to existing information, such as the notes people leave in the margin of a book. But annotations are frequently digital, like the tags people add to Facebook photos to make searching through albums easier. New augmented-reality technology now adds digital notes to physical objects, such as additional nutritional information on food. Annotations are perhaps more akin to the distinct scratches and scents animals use to mark territory [8].

We sometimes view digital annotations in our territories with an augmented-reality device. When we view street information with such a device, the annotations are overlaid on an image of the street as seen by the device. And when sports fans watch a game at home, they view annotations registered on individual players as they move through video frames. The street image and sports video both serve as surfaces for annotation display.

Presenting digital annotations on annotation surfaces poses a number of technical problems. Annotations must be aligned both spatially and temporally - we expect digital scents added to cars to also travel and remain aligned despite augmented-reality device orientation. And if the devices are to become popular, how should users create and align annotations efficiently and accurately?

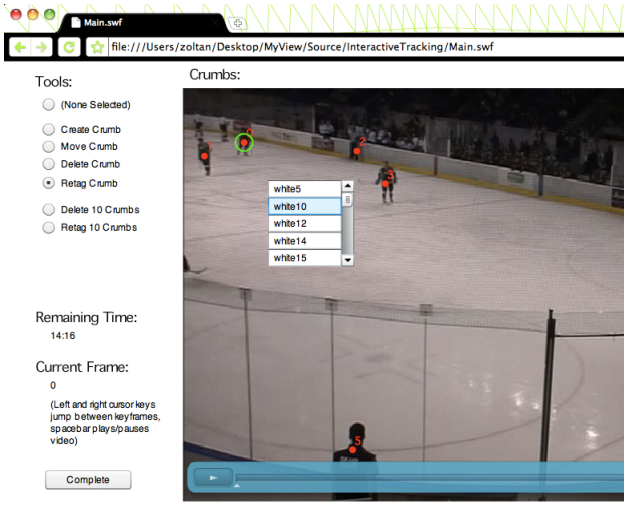

Figure 1. Novel "breadcrumb" interface to assist users annotating moving objects with video

Current tools to create annotations often rely on tedious manual registration of information on images or video. YouTube users currently align annotations with a computer mouse for each frame of their videos (Figure 2). But research tools, such as the Video Performance Evaluation Resource (VIPER) [6], are beginning to provide rudimentary features to propagate annotations and maintain correct registration of annotations on objects. The VIPER features are the first steps to computer assistance to leave our annotation marks.

In this paper, we present a novel interaction technique, known as "breadcrumbs", that provide advanced assistance to users creating annotations for moving ice hockey players. We employ computer vision techniques to track objects and simplify the task of aligning annotations of individual players in video frames. Users do not need any computer vision expertise to create the annotations. We measure the accuracy of operators and also the time they spend annotating, and compare their performance when they are given no computer assistance.

Submitted for review to CHI 2011.

But we also provide a cautionary tale to researchers creating annotation systems. Our prototype system frequently created more work for operators than it saved. So although operators have created accurate annotations with our breadcrumbs interaction, we could not detect any significant advantage of computer assistance. We warn you that computer annotation assistance may disrupt your concentration.

Even though our system currently only provides assistance to position annotations, and our footage was limited to sports video, our results are more broadly relevant. We will frequently need to add our digital paw-prints to objects moving on augmented-reality device screens. And surveillance or human behaviour researchers should be aware that computer assistance for annotation has not yet been shown to improve operator performance.

Future work on computer assistance will combine operator strengths handling uncertain conditions with the speed of computer vision techniques. But operators should be provided with control over the computer assistance so that they can avoid disruption. Research has yet to provide substantive evidence that computers can assist users creating annotations. But we believe that computer assistance will yet make creating annotations as easy as taking a whizz on a mountain trail.

REVIEW OF PREVIOUS WORK

Manual Annotation

Researchers have developed a variety of software tools for manually annotating objects in image and video, such as the crowd-sourced LabelMeVideo [17], or the Video Performance Evaluation Resource (VIPER) [6]. Although manual annotation tools take advantage of the intelligence and innate visual processing of operators, they are painfully slow to produce annotations, typically requiring about ten minutes of operator time for every minute of video footage. Distributing tracking work to operators is expensive, and the costs grow in step with the length of video.

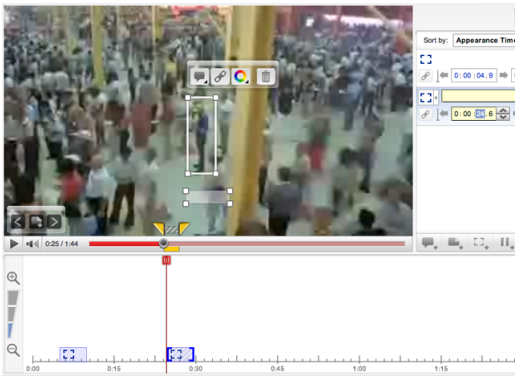

Figure 2. YouTube interface to position annotations on video frames

Operators use a variety of "common sense" techniques to recognise and track objects. Humans can deduce the relative heights of objects to distinguish between short and tall objects. Symbols or patterns of color on objects also help us to recognize and track objects. And from childhood, we develop intuitions for trajectories to help us follow objects when they are briefly occluded.

But human annotation is slow. Operators must perform precise motor movements to indicate the positions of objects. Humans attention limits the number of objects simultaneously tracked. Operators typically track a single object through a sequence of video frames, returning to the same frames to track other objects later.

Interaction techniques have been developed to improve the performance of manual annotation tools. State of the art tools, such as VIPER [6], incorporate interpolation and propagation features to allow operators to skip video frames when annotating. Interaction techniques such as "Click to Pause" [9], also reduce the number of operations during video annotation.

So although human annotation is slow, manual tools are readily available. Sports coaches can use annotation tools such as Sporideo STEVA [1] to determine player statistics such as time spent in particular zones. IBM developed a general-purpose video annotation tool, VideoAnn [13], that researchers could use to embed annotations in MPEG7 metadata. Manual annotation is necessary to deal with the huge variety of image and video to be annotated, with applications for sports, surveillance or market research.

Automated Annotation

Computer vision is a broad term for computing techniques that extract scene and feature information from video sequences. Computer vision researchers have developed computer algorithms to detect and track objects in images and video [16]. Some may think that algorithms are the only way to annotate both enormous bodies of existing video and the growing stream of new video. But the algorithms must be tuned by experts to specific applications. The techniques struggle with many object conditions, such as when captured by moving cameras, shot in challenging lighting conditions or when objects overlap each other.

Computer vision techniques must be tuned before they can detect and track objects. Simple techniques are tuned to detect foreground objects as blobs of pixels by comparing images with a known background image. But since this technique is only appropriate with stationary cameras, more recent object detectors are tuned to recognize objects based on image patterns.

But even after detection tuning, the techniques frequently fail to detect objects, or detect objects that should not be tracked. Objects that pass close to each other are sometimes mistaken for a single object. If an object is partially obscured then the detection also fails. Shadows are sometimes mistaken for objects that should be tracked. The vision techniques struggle to recognise objects too. When an object leaves a scene and subsequently reenters, the techniques often detect and track the object as two distinct objects.

Other technologies are available for object tracking, such as microwave frequency tags [4]. But the systems require tracked objects to have tags attached, complicating system setup and raising system costs. And because some situations have sufficiently controlled video conditions, completely automatic annotation systems, such as the AMISCO soccer player tracker [4], are in widespread use today. Researchers currently develop analysis tools that assume accurate annotation of stationary indoor security footage, such as the sentient environment developed at the Sarnoff Corporation [18].

Hybrid Annotation: Computer Assistance for Manual Annotation

Hybrid annotation is a form of shared control between human and machines. Control is the use of feedback to allow a system to deal with uncertain operating conditions. Shared control is well demonstrated by the interactions of pilots and their aircraft autopilot. The autopilot is expert at steadily navigating situations of reduced visibility but can't compete with the human ability to deal with novel situations.

In the 1960s, Licklider envisioned shared control as a productive combination of goal-oriented operators and rapid execution of routine operations by machines [10]. In the years that followed, humans became supervisors of complex systems, their decisions integrated with feedback loops to cope with a range of uncertain operating conditions. As Sheridan highlighted, unpredictable jobs cannot be easily automated and humans cooperation with machines will remain necessary [12].

The uncertainty of video content suggests that humans should share control with machines to efficiently track objects. Humans can intelligently deploy appropriate machine vision techniques to novel video conditions. Annotation time will include machine processing time to track objects, and operator time to supervise the accuracy of the machine processes.

Researchers at the University of Washington applied shared control to animation, interpolating operator outlines in video with computer vision techniques, although they did not report any measures of operator performance [3]. Ivanov developed an interactive system for operators to combine annotations from different tracking sources, although he assumed that highly accurate computer vision estimates would be readily available and also neglected to assess system performance [2].

DeCamp showed how operators could merge annotation "tracklets" from realtime computer estimates with the accuracy of unassisted tools, but in an order of magnitude less operator time [5]. Although, Decamp performed his evaluation with only two experimental subjects (one a computer vision researcher), his approach is promising. And recently, Vondrick developed a crowd-sourced annotation system that assisted video annotators with interpolations of object locations [15]. He showed how in simple video situations, accurate annotations are created at reduced operator costs by providing computer assistance.

For typical annotation tasks, operators will not have advanced computer vision expertise. For instance, a laboratory gathering annotations for analysis may assign video annotation tasks to non-experts. In this paper, we will present a design for a hybrid annotation system to assist non-expert operators annotating video. We also measure the performance of the operators and show that performance of annotation systems that provide computer assistance is crucially related to the quality of the assistance provided.

HYBRID ANNOTATION INTERFACE DESIGN

Design Requirements

In this paper, we consider the task of creating annotations of position for moving objects. We consider a situation where movements of objects are captured with a video camera. In our situation it is sufficient to locate objects within the bounds of the video frames. Given this situation, we developed the following design requirements for an interface to provide assistance creating the annotations:

- DR1: Operators should be able to perform three basic supervisory operations: accepting, rejecting or modifying computer estimates. Computer estimates track erroneous objects, known as false positives and the operator should be able reject such errors. Computer estimates often fail to track objects between occlusions of the objects, and operators should be able to modify the estimates to indicate when an object path is a continuation of a previous path. Computer estimates also fail to detect objects, perhaps because the color of the object is not distinct against background objects, and the operator should be able to add missing annotations.
- DR2: Operators without computer vision expertise should be able to create annotations efficiently and need minimal training to reach efficient annotation rates.
- DR3: Operator performance should be faster than manual annotation systems. Computer assistance should target common tasks to improve overall system performance.

We excluded some hybrid annotation issues from our work:

- Our interface will not recognise objects, but only track objects between frames with unique identifiers.
- Our interface will not allow operators to adjust properties of computer vision tracking algorithms, and will just rely on offline computer vision tracking results.
- Our interface will only allow operators to view a single video angle.

Design Features

Paper prototypes of annotation interfaces were discussed with two computer science students and we developed a high fidelity prototype from the discussions. We asked four other computer science students, ranging in age from 23 to 30 years old, to correct annotations of ice-hockey player positions for a single video frame with the high fidelity prototype. From our observations and interviews with the stu-

dents, we developed a model of annotation that we call the "breadcrumb" model.

Users are encouraged to think of annotation points as like crumbs, dropped to mark a path by Hansel and Gretel in the classic Brothers Grimm fairy tale [7]. Objects to annotated with a position within a video frame are equivalent to the individuals Hansel or Gretel. The crumbs that indicate the path of an object, such as Gretel, are all marked with the same "Gretel" tag. Our interface implements four standard functions to edit the annotations for a single frame:

- Create Crumb: A crumb is created at any mouse click position, perhaps similar to dropping a crumb in a fairy tale forest.
- Move Crumb: When the user clicks, the crumb nearest to the mouse pointer is moved to the mouse pointer location, and if the user drags, the crumb can be relocated further.
- Delete Crumb: When the user clicks, the crumb nearest the mouse pointer is removed.
- Retag Crumb: When the user clicks, a scrolling menu appears at the mouse pointer location, and if the user selects a menu item, the nearest crumb to the mouse pointer location is tagged with the menu choice. Menus contain a set of tags in a predictable order that are not already used by other crumbs. This function allows operators to specify the object that the crumb annotates with a position.

We also implement a feature to generate "interpolated crumbs" between explicitly annotated object positions. These implicit annotations are colored blue to distinguish them from red explicit annotations. The interface maintains a list of the tags applied to crumbs in all video frames - the list is equivalent to an object list (i.e. Hansel, Gretel) tracked in the video. If a frame does not contain an explicit crumb for an object annotated in any other frame, then the interface examines neighboring frames for explicit crumbs and generates an uneditable implicit crumb. The position of the implicit crumb is function of the time difference between the displayed frame and the frames containing explicit frames - similar to the "tweening" animation technique [11]. These interpolated annotations result in piece-wise linear rendering of annotations and allow users to reduce the number of frames they annotate explicitly.

Standard video play and pause button and draggable playhead are provided and the left and right arrow keys jump forward or backward between video keyframes.

The design provides features for editing batches of annotations spread across a number of frames, suitable for correcting computer vision generated crumbs. Batch operations operate on a limited number of ten crumbs to avoid editing crumbs that could be correct in much later frames.

- Delete 10 Crumbs: When the user clicks, the crumb nearest the mouse pointer is deleted. But the interface also deletes up to ten crumbs in subsequent frames with the same tag as the nearest crumb.
- Retag 10 Crumbs: When the user clicks, and if the user selects a menu item, the nearest crumb to the mouse pointer location is tagged with the menu choice. But the interface also retags up to ten crumbs in subsequent frames with the same original tag.

A file of crumbs representing annotation positions can be loaded by the interface, such as positions generated by a computer vision tracking algorithm.

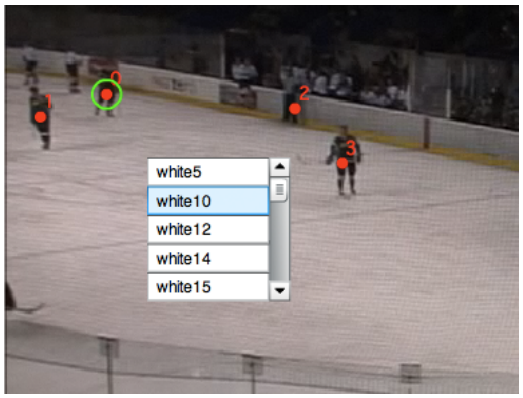

Figure 3. Computer suggested crumbs on objects captured with video. The displayed menu is to choose a new tag for the crumb highlighted in green. Note that the referee has a crumb suggestion with temporary tag "2" - the user will remove the crumb since they only want to annotate players. Also note two white players in the frame background have not been given crumbs.

The breadcrumb metaphor was expected to reduce training times, required by DR2. The metaphor will be familiar to many people and introduces the concepts that the operator will have to grasp to perform the position annotation task. All the crumb operations described combine to satisfy DR1. And by providing computer vision suggested crumbs in addition to state-of-the-art features such as interpolation, we hoped the interface would outperform a manual annotation system and satisfy DR3.

To our knowledge, no researchers have used a breadcrumb metaphor for interfaces that create annotations of object positions. And although computer vision researchers have examined how to share the work of object tracking between computers and operators, we are the first researchers to consider how computer assistance can help non-computer-vision-experts to perform annotation tasks. To evaluate our design and measure the performance of operators, we designed a controlled experiment. As will be described below, we found that the success of our design is related to the quality of the computer vision suggested crumbs.

CONCEPT EVALUATION

Experiment Tasks

Participants annotated keyframes at every sixth frame for two video sequences, Angle A and Angle B, of a college ice hockey game. The sequences were from the same period of the game, but participants were unaware of this and they could view only one video during the experiment, so we assumed that participants treated the sequences as from

different periods. The video was recorded from cameras that were static for the duration of the sequences. Player to be tracked in the video did not suffer from foreshortening.

Figure 4. Experiment workstation

Participants completed a training session with video from a training angle different to Angle A or Angle B. Participants had to complete training tasks for both conditions before progressing to any timed trials. Written instructions introduced the concept of annotation as similar to dropping breadcrumbs to indicate a path of movement. To complete training for the assisted condition, three frames with computer vision suggestions for crumb positions had to be annotated to the satisfaction of the experimenter. To complete the manual condition training, a single hockey player had to be annotated for twenty frames. The participants tried the delete, retag and create crumb operations under instruction from the experimenter during the assisted condition training. After the basic operations were understood, the experimenter introduced the "retag many" and "delete many" operations. The interpolation feature was introduced during the manual condition training.

Participants then each completed two fifteen minute trials - the "assisted" and "manual" trials. The order of the trials were counterbalanced so that half the participants conducted the manual trial first, the other participants conducted the assisted trial first. The camera angles was also balanced so that participants were randomly assigned to one of four conditions:

- Cam A Assisted, Cam B Manual
- Cam B Assisted, Cam A Manual
- Cam A Manual, Cam B Assisted
- Cam B Manual, Cam A Assisted

Participants were provided with a list of tags corresponding to the numbers on the ice-hockey player shirt. They were told to choose tags for hockey player crumbs as accurately as possible, but that consistency was more important than getting the tag correct. Participants were instructed to only track players and not to track referees, players "off the ice", the goalie nor any spectators.

Participants could play or pause the video and move between keyframes with the keyboard left and right cursor keys. The video playhead could also be dragged to move between keyframes. The time remaining was displayed for the trial and the annotation interface disappeared as soon as the trial ended.

The suggested annotations were generated by a simple computer vision algorithm prior to the experiment and participants received the same suggestions. Detection and tracking algorithms subtracted a background image, masked non-rink areas, detected blobs within a tuned range of sizes, and tracked blobs with a color-based particle filter. The suggested annotations were stored in an XML format and loaded into the annotation tool at the start of the assisted conditions. Blob size restrictions were chosen to produce an acceptable detection rate.

Equipment

Each participant used a 17 inch widescreen monitor, and a standard mouse pointer. The video was presented at a resolution of 854 x 480 pixels. Participants could play the video at 29.97 frames per second.

Measurements

During the experiment, we measured the time each operator spent performing seven operations - Create Crumb, Move Crumb, Delete Crumb, Retag Crumb, Delete 10 Crumbs, Retag 10 Crumbs and No Operation. Before performing an operation, participants had to reselect an operation by mouse from a menu at a consistent location. No keyboard shortcuts were provided. By forcing participants to reselect the operation from the menu, we expected operation times to be more consistent.

After each trial, we recorded participant's annotations in an XML format for analysis. We asked participants to complete the NASA Task Load Index weighting and component surveys immediately after each condition [**?**]. At the end of both trials, participants were asked to complete an online survey to record their age, computer experience, reported strategy and preference for each condition.

Participants

We recruited participants from internet personals site Craigslist and the University of British Columbia Institute for Computing, Information and Cognitive Systems experiment recruitment system. In addition to snacks, drinks and a public transit ticket, participants were offered $10 for an hour of their time. Participants were also told that they would be awarded $30 for the most accurate and complete annotations. They were required to have basic English reading and writing skills, to be able to use a standard computer monitor and mouse, and to see a normal range of colors.

Of the twenty participants that took part in the experiment, twelve were women. Participants were aged from 17 to 57, with an average age of 28. All had previous experience using computers recreationally or for employment. Although two participants were experienced software developers, none of the participants reported knowledge of computer vision.

RESULTS

(a) Training Angle

(b) Trial Angle A

(c) Trial Angle B

Figure 5. Example video for annotation.

Timing Measurements

In the fifteen minutes that participants spent on the manual condition, 50% of participants annotated less than 426 frames, or about 14 seconds of footage. In the same amount of time on the assisted condition, 50% of participants annotated less than 342 frames. These values were calculated by looking for the last frame annotated in the manual condition, and the last frame different from the computer vision suggestions for the assisted condition. Values of "frame-reached" are plotted for each participant in Figure 6. Note that participants may not have completed annotation in any of the frames.

Analysis of variance within participant measures of frame-reached revealed a significant effect of condition ($F_{1,19} = 6.69$, $p = 0.02$), but we could not determine if there was a significant effect of order between participants ($F_{1,19} = 1.28$, $p = 0.27$), nor if there was a significant interaction between order and condition ($F_{1,19} = 0.07$, $p = 0.79$).

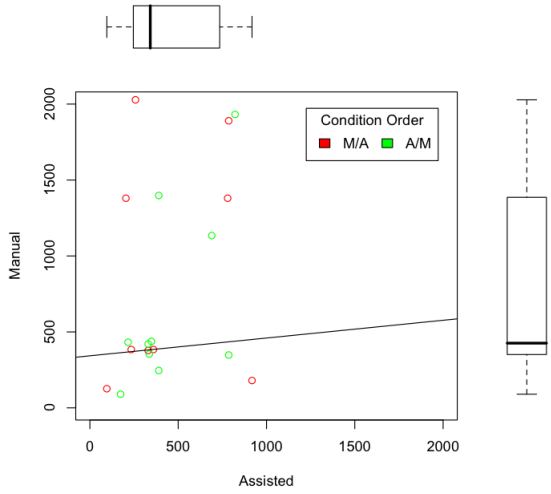

Figure 6. Plot of frame reached in both conditions for each participant.

The time that participants spent on each annotation operation is plotted in Figure 7. During the assisted condition, 50% of participants spent more than 20% their time deleting annotations and a similar number of participants spent more than 25% of their time retagging annotations. In the manual condition, 50% of participants spent more than half of their time creating annotation points but another 50% of participants spent less than 15% of their time tagging the annotations they created.

Analysis of variance within participants measures of operation time revealed a significant effect of operation type ($F_{4,19} = 93.71$, p ¡ 0.01) but we could not determine a significant effect of condition ($F_{1,19} = 0.12$). Analysis of variance also revealed a significant interaction effect of condition and operation within participants ($F_{4,19} = 87.54$, p ¡ 0.01).

In a situation when computer vision was giving good assistance to operators, we would expect to see that operators spent more time retagging annotations and less creating and

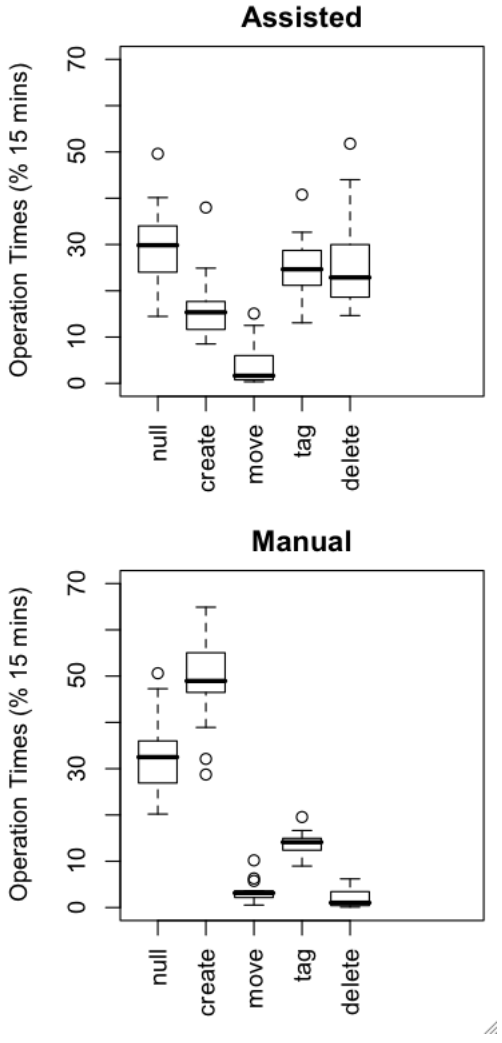

Figure 7. The proportion of operator time spent on five different operations (null is a catch-all category for time not spent on any operation)

deleting annotations.

Accuracy Measurements

To calculate measures of annotation accuracy, we performed two data processing steps. A "ground truth" of accurate annotations for both processing steps was generated by hand.

The first step normalized participant tags, since we were interested in how consistently participants tracked players - not in their ability to recognize specific player numbers. Annotations in the first frame of each participants data were paired to annotations in the ground truth using a geometric distance measure and each participants annotations were then retagged. The pairing was only performed for the players in scene for the first frame and did not account for players that subsequently entered the scene. Most operators did not annotate past situations when new players entered the scene.

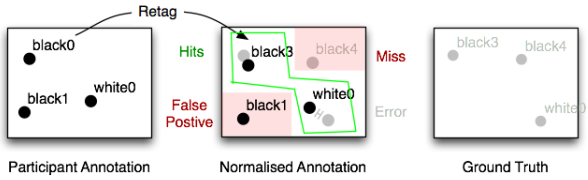

Figure 8. Normalization retag example, and hit, miss and false-Positive detection

The second processing step compared annotations for each frame based on annotation tags. For each frame in the ground truth, annotations from an equivalent frame for each participants data were retrieved, including interpolated annotations. Annotations that shared the same name were considered hits and an error value was calculated based on the geometric distance between the ground and participant annotations. To compare the performance of the assisted and manual conditions, we calculated average hit rates, miss rates, false-positive rates and error values for each frame (Figure 8).

Note that if the normalization step failed to pair a participant annotation with a ground annotation, then we recorded both a false positive and a miss. And it should be noted that even though we used two different camera angles, the number of objects in the keyframes during the experiment was not uniform. Other studies of annotation performance will find different ratios for "frames annotated to operator time".

We developed a metric to calculate each participants accuracy. The metric accounts for misses, false-positives and hits for each frame, but does not account for mis-tags, or the error value of hits. For each participant, we sum the number of hits for each keyframe, up to the last keyframe they annotated. We then subtract the number of misses and false positives over the period. The result is equivalent to the area illustrated in Figure 10, and we refer to the measure as the "annotation work".

The maximum annotation work during the same period is equivalent to the maximum number of hits during the period

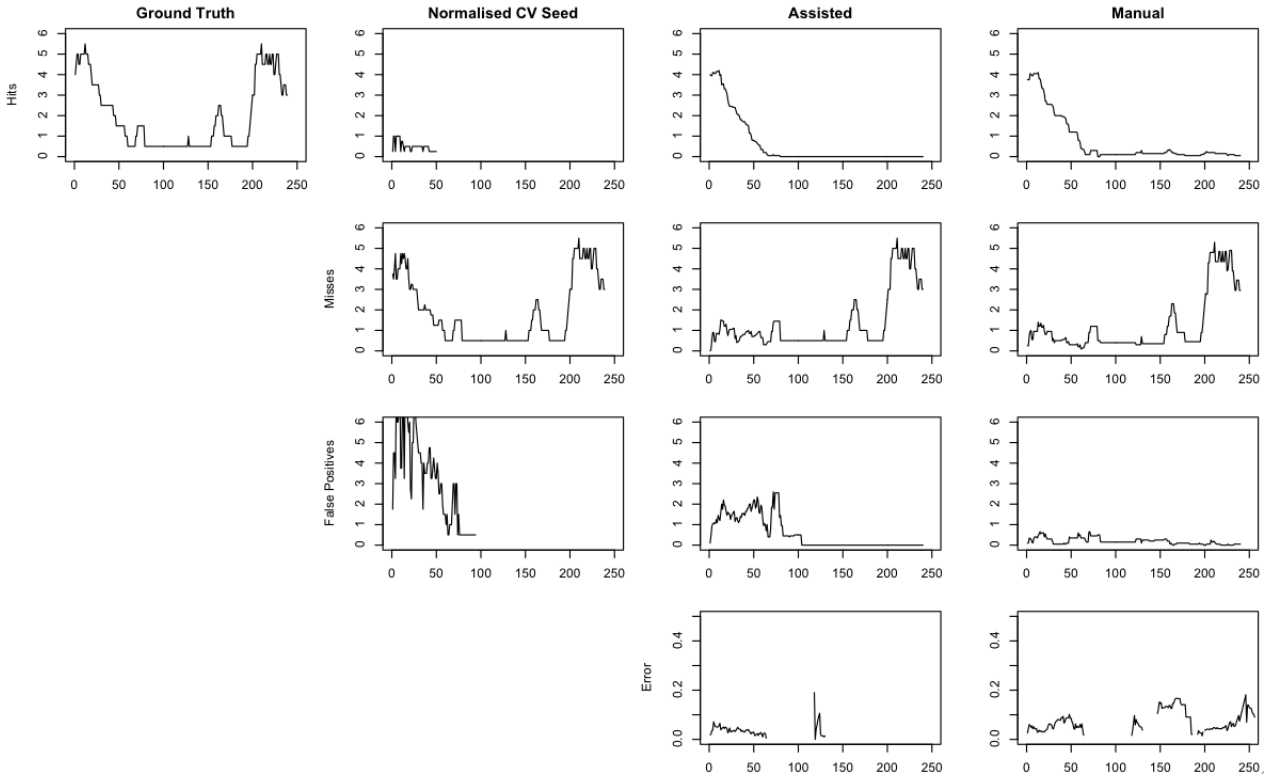

Figure 9. Accuracy measures for the first 250 keyframes

illustrated in Figure 10. We divide each participants annotation work by this maximum annotation work. This result of the calculation will range from an unbounded negative number to one - from participants with many false positives to participants with a perfect hit count over their annotation period.

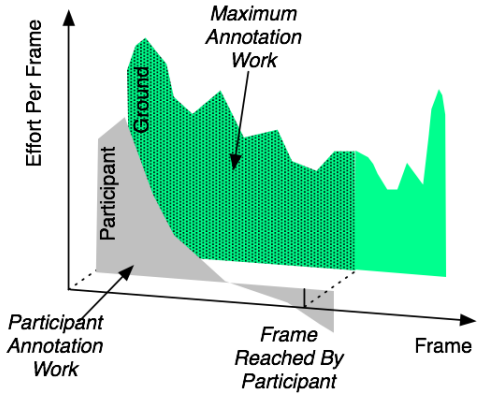

Figure 10. Regions for participant accuracy calculations

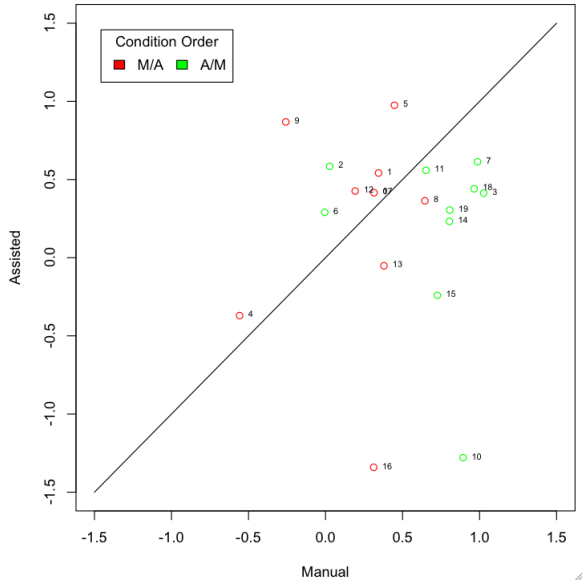

Figure 11. Distribution of accuracy measures for each participant

Accuracy measures for each participant are presented in Figure 11. Analysis of variance within the participants does not reveal any effect of condition on their accuracy ($F_{1,19} = 1.94$, $p = 0.18$). And despite the appearance of an effect of order in the graph of accuracies, we could not determine a significant effect of order between participants ($F_{1,19} = 3.67$, $p = 0.07$), nor a significant interaction between order and condition ($F_{1,19} = 0.70$, $p = 0.41$).

Subjective Measures

When asked "What condition (Manual or Assisted) did you prefer and why?", only five participants preferred the assisted condition. Participants reported annoyance with the computer vision tracking mistakes. One participant explained how mistakes in the assistance interrupted their "flow" of work. Some participants explained that they felt more in control of the manual condition and didn't have to worry about erroneous annotations cropping up unexpectedly. Some of those who did prefer the assisted condition felt that the computer "did more work" than the manual condition.

When asked "What strategy did you use to annotate the players?", participants reported techniques for tracking players from frame to frame, taking advantage of the interpolation or computer vision features, and following individual players. Six participants mentioned memorising the player tags or positions so they didn't have to flick back to previous frames. Two participants reported looking for players travelled in straight lines to take advantage of the interpolation feature. One participant reported looking for events the computer vision struggled with, such as players leaving the scene. Six participants explicitly described focusing on a single player and how they returned to early frames to find new players to track.

We collected reliable TLX weight measures for only eight participants, although all four conditions of camera angle and assistance order were equally covered. The values presented in Table 1 are for both the TLX are normally measured for eight participants and for an "un-individuated" measure based on average task weights for all twenty participants. But although we conducted an analysis of variance among the eight accurate TLX measurements, we could not determine if condition was a significant factor in task load differences among the eight participants ($F_{1,7} = 0.25$, $p = 0.63$), nor if order had a significant effect ($F_{1,7} = 0.49$, $p = 0.51$), nor a significant interaction effect of order and condition ($F_{1,7} = 0.28$, $p = 0.62$).

Table 1. Average reported NASA TLX (standard deviation in parenthesis) (0: Low/Good, 100: High/Poor)

	Individual Weights	Average Weights
Assisted	30.32 (12.71)	23.61 (7.01)
Manual	31.94 (13.58)	32.45 (10.57)

DISCUSSION

Our design aimed to improve the performance of operators annotating moving objects in video. We aimed to create a design so that operators without computer vision expertise could rapidly become proficient creating annotations. The design should inform research on computer assistance for a broader range of annotation tasks for a range of applications such as surveillance and augmented reality.

We could not detect any significant effect of computer vision assistance on the accuracy of operators annotating video. Worryingly, operators made significantly more progress on annotations in the manual condition. And although we could not detect any significant difference among operator task load reports, operators did complain that computer vision assistance interrupted them. This may have been because of shortcomings with the training procedure, because of unrealistic experiment design or because of failings of our interface design.

We did not measure the performance of each participant to assess their skill level immediately after training. We observed some experiment participants struggling with the interface because they had not fully understood the interface. The consistency of our training is unclear and may have benefitted one condition more than the other. The oversight to measure the effects of the participant training also complicates estimates of how operators learned to use the tool. And statistical tests failed to find a significant learning effect on the measures of accuracy and frame-reached within subjects.

During the experiment, participants were forced to reselect tools after each operation. Such a restriction is unlikely in real-world applications. Computer vision suggestions had to be deleted from many frames, so operators were bogged down reselecting the delete tool. In the manual condition, operators could use interpolation to reduce the number of operations per frame and suffered less from tool selection delays.

An improved interaction design would allow the operator to adjust properties of the computer vision assistance and reduce the time spent on delete operations. For example, if operators could edit the mask used to exclude image areas from computer vision analysis, then they could fine tune the mask to remove the goalie from computer vision suggestions. "Onion skins" to overlay details from previous frames could also prevent operators from flicking back to check annotation details.

In a broader range of annotation tasks, such as adding names or ratings to objects, operators may improve the quality of computer assistance by tuning. In the case of computer vision, operators could create tracking masks and perhaps inspect computer vision detection blobs. Future work should examine how to expose these tuning properties to non-expert users.

Research on other annotation domains should consider the false positive rate of computer vision assistance. Two types of false positives demanded operator time for corrections - frequent objects, such as the referees and goalies in our experiment, and infrequent objects, such as bystanders. Interaction techniques that target frequent false positives should have the greatest impact on operator performance.

Future annotation research could also develop a performance metric that accounts for operator time to confirm that a scene has no objects to track. Future work should also consider other metrics of accuracy such as the Multiple Object Tracking metrics developed by the Workshop on Classification of Events, Activities and Relationships [14].

CONCLUSION

We expected that computer vision assistance would provide a clear improvement on object annotation tasks. But our experiment is a warning to researchers that computer assistance does not always save time.

We reviewed previous work on manual and automated annotation. To our knowledge, researchers have yet to measure how effectively annotators are assisted in their tasks by automated techniques. We also suggest that research has not yet stated that annotation systems such be designed for unskilled operators.

We presented our design to improve the performance of operators annotating the positions of objects. We developed a novel breadcrumb metaphor with an iterative design process to satisfy requirements for an annotation tool for unskilled operators.

We evaluated our design with a controlled user study to annotate ice-hockey players in video. Although the computer assistance we provided was to the detriment of operator performance, we think our results provide evidence for researchers working on a range of annotation tasks and applications that operators should be given control over computer assistance.

Researchers have yet to provide evidence that operator performance can be improved by computer assistance. But the success of shared control systems, that combine operator strengths handling uncertain conditions with the speed of computer techniques, suggests that research will soon find the best ways to improve annotation.

REFERENCES

1. STEVA Sports Software Inc. `http://sporideo.com/home/default.aspx`.
2. Y. I. A, E. S. B, C. W. A, and I. K. C. Tracking people in mixed modality systems.
3. A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz. Keyframe-based tracking for rotoscoping and animation. *International Conference on Computer Graphics and Interactive Techniques*, 23(3):584, 2004.
4. S. Barris and C. Button. A review of vision-based motion analysis in sport. *Sports medicine (Auckland, N.Z.)*, 38(1212):1025–43, 2008.
5. P. DeCamp and D. Roy. A human-machine collaborative approach to tracking human movement in multi-camera video. *Conference On Image And Video Retrieval*, 2009.
6. D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. pages 167–170, 2000.
7. J. Grimm and G. W. *The Complete Grimm's Fairy Tales*. New York: Pantheon Books, 1944.
8. L. F. G. A. B. B. Hemker, T. P. Population characteristics and movement patterns of cougars in southern utah.
9. M. V. Ilich. Moving target selection in interactive video. *UBC Information Repository*, 2009.
10. J. C. R. Licklider. Man-computer symbiosis. *Human Factors*, (March):4–11, 1960.
11. W. T. Reeves. Inbetweening for computer animation utilizing moving point constraints. *International Conference on Computer Graphics and Interactive Techniques*, 15(3):263, 1981.
12. T. B. Sheridan. *Telerobotics, automation, and human supervisory control*. MIT Press, 1992.
13. J. R. Smith. Visual annotation tool for multimedia content description. *Proceedings of SPIE*, 4210:49–59, 2000.
14. B. K. B. R. R. R. T. M. M. Stiefelhagen, R. and J. Garofolo. The clear 2007 evaluation. *Springer-Verlag, Berlin, Heidelberg*, 2008.
15. R. D. Vondrick, C. and D. Patterson. Efficiently scaling up video annotation with crowdsourced marketplaces. *European Conference on Computer Vision (ECCV) Crete, Greece*, 2010.
16. A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4), 2006.
17. J. Yuen, B. Russell, C. Liu, and A. Torralba. Labelme video: building a video database with human annotations. *ICCV, Kyoto, Japan*, 2, 2009.
18. T. Zhao, M. Aggarwal, T. Germano, I. Roth, A. Knowles, R. Kumar, H. Sawhney, and S. Samarasekera. Toward a sentient environment : real-time wide area multiple human tracking with identities. *Machine Vision and Applications*, pages 301–314, 2008.