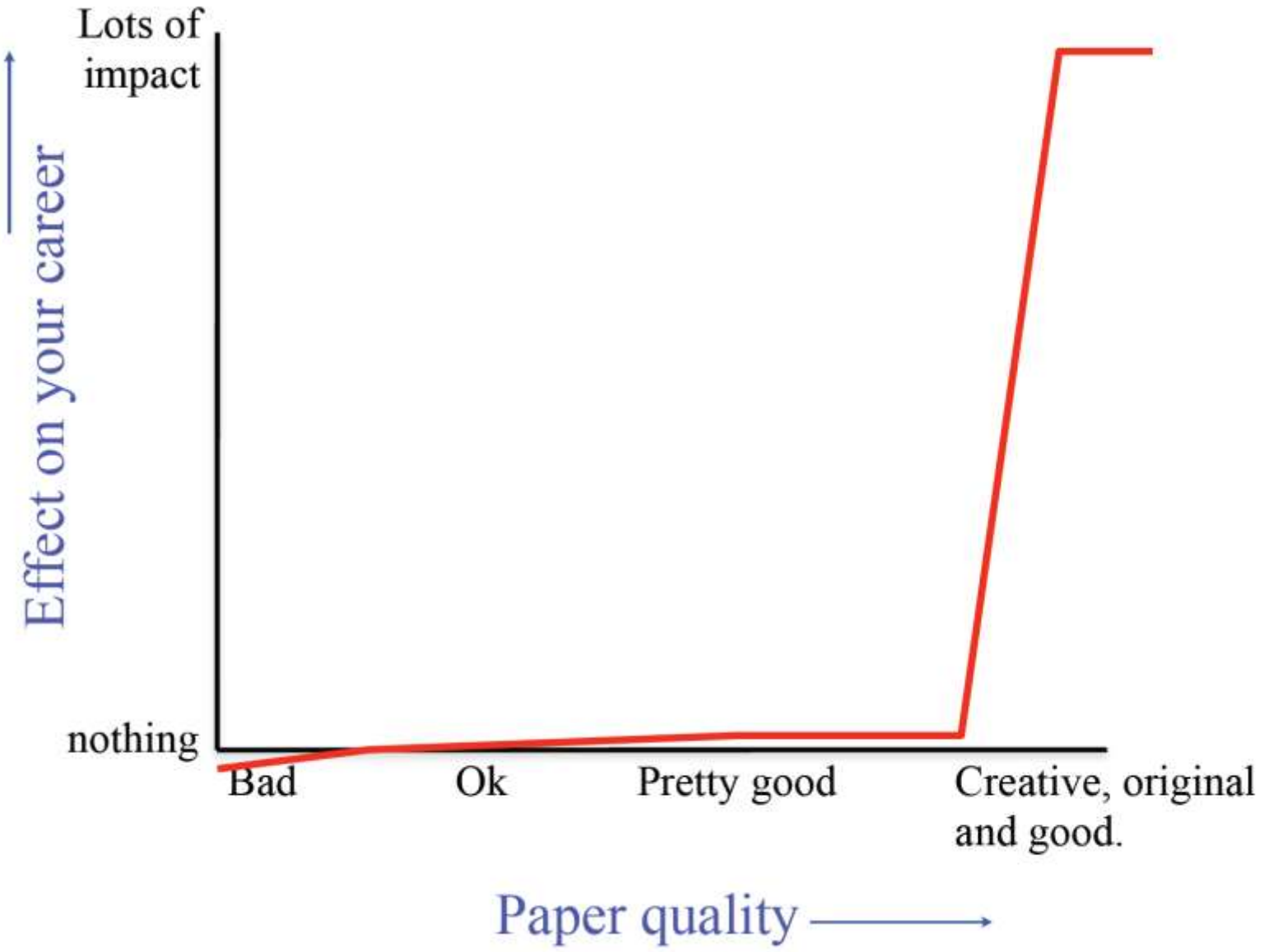
Paper Writing I: Introduction

Disclaimer: I am not a good writer.



Bill Freeman's note: How to write a good CVPR paper
https://billf.mit.edu/sites/default/files/documents/cvprPapers.pdf

Kajiya description of what reviewers look for.

The most dangerous mistake you can make when writing your paper is assuming that the reviewer will understand the point of your paper. The complaint is often heard that the reviewer did not understand what an author was trying to say

Bill Freeman's note: How to write a good CVPR paper
https://billf.mit.edu/sites/default/files/documents/cvprPapers.pdf

You must make your paper easy to read. You've got to make it easy for anyone to tell what your paper is about, what problem it solves, why the problem is interesting, what is really new in your paper (and what isn't), why it's so neat.

Bill Freeman's note: How to write a good CVPR paper
https://billf.mit.edu/sites/default/files/documents/cvprPapers.pdf

Organization

1. Abstract
2. Introduction
3. Related Work
4. Method
5. Data and analysis (optional)
6. Results
    1. Quantitative result
    2. Qualitative result
7. Conclusion/discussion

Social Saliency Prediction

Hyun Soo Park Jianbo Shi
University of Pennsylvania
{hypar, jshi}@seas.upenn.edu

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect

sion solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

While there are many factors involved in joint attention,

Park and Shi, “Social Saliency Prediction”, CVPR 2015


Social saliency
Social saliency
Camera
Top view


London 2012
FIBA
London 2012
London 20

1. Abstract
2. Introduction

Kajiya: write a dynamite introduction

How can you protect yourself against these mistakes? You must make your paper easy to read. You've got to make it easy for anyone to tell what your paper is about, what problem it solves, why the problem is interesting, what is really new in your paper (and what isn't), why it's so neat. And you must do it up front. In other words, you must write a dynamite introduction. In your introduction you can address most of the points we talked about in the last section. If you do it clearly and succinctly, you set the proper context for understanding the rest of your paper. Only then should you go about describing what you've done.

Bill Freeman's note: How to write a good CVPR paper
https://billf.mit.edu/sites/default/files/documents/cvprPapers.pdf

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph

- In this paper, …

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph

- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph

- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph

- Problem definition: input / output

Abstract

This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.

1. Introduction

Imagine an artificial agent such as a service robot operating in a social scene as shown in Figure 1. It would detect obstacles such as humans in the scene and plan its trajectory to avoid collisions with the obstacles. It may plan a trajectory that passes through the empty space between the audiences and performer. This trajectory intrudes on the social space created by their interactions, e.g., occluding the sight of the audiences, and thus, it is socially inappropriate. We expect the artificial agent to respect our social space although the boundary of the social space does not physically exist. This requires social intelligence [31]—an ability to perceive, model, and predict social behaviors—to be integrated into its functionality.

Joint attention is the primary basis of social intelligence as it serves as a medium of social interactions; we interact with others *via* joint attention[1]. Understanding joint attention, specifically knowing where it is and knowing how it moves, provides a strong cue to analyze and recognize group behaviors. It has been recognized that computer vision solutions can provide a large scale measurement for developing a computational representation of joint attention. The challenges are: (1) human detection and tracking failure in the presence of occlusions, (2) scene variability, e.g., the different number, scale, and orientation of social groups, and (3) inaccurate measurements of gaze directions.

[1]Gaze directions are correlated with joint attention in quasi-static social interactions while motion becomes a dominant factor in rapid dynamic interactions as shown in Figure 3(b).

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

While there are many factors involved in joint attention, our main question is: can one predict joint attention using social formation information alone, without the gaze information of each member?

In this paper, we show that it is possible to empirically learn the likelihood of joint attention called *social saliency* as a function of a social formation, a spatial distribution of social members, using data from first person cameras. Three key properties of our predictive joint attention model are: a) it is scale and orientation invariant to social formations; b) it is invariant to scene context, both indoors and outdoors; c) it is robust to missing data. Once this model is constructed, a sparse point cloud representation of humans can be used to predict the locations of joint attention as shown in the inset image of Figure 1, without any directional measurement such as gaze directions—we measure and learn this predictive model in the first person view, and apply in the third person view.

To construct such joint attention model, we use first person cameras. With multiple first person cameras, joint attention can be precisely measured in 3D since the ego-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph

- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph

- Problem definition: input / output
- Brief presentation of results

motion of the cameras follows gaze behaviors of the wearers [2, 9, 23]. Furthermore, we can simultaneously compute the 3D positions of the wearers to provide precise measurements of social formations. These first person in-situ computational measurements of the geometric relationship between joint attention and social formation can be applied in a variety of third person social interaction scenes including a basketball game where the players strategically take advantage of spatial formations.

Contributions To our best knowledge, this is the first work that provides a predictive model encoding the geometric relationship between joint attention and social formation, using in-situ 3D measurements from first person cameras. This paper presents three core technical contributions: (1) a construction of a social formation feature that is scale and orientation invariant inspired by electric dipole moments; (2) a method of discovering multiple social groups using scale space extrema in a spatial distribution of social members; (3) a consolidation of social interaction data reconstructed in difference scenes, which allow us to learn and infer social saliency in a unified coordinate system. Our method can predict social saliency from a third person video or image of social interactions.

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

motion of the cameras follows gaze behaviors of the wearers [2, 9, 23]. Furthermore, we can simultaneously compute the 3D positions of the wearers to provide precise measurements of social formations. These first person in-situ computational measurements of the geometric relationship between joint attention and social formation can be applied in a variety of third person social interaction scenes including a basketball game where the players strategically take advantage of spatial formations.

Contributions To our best knowledge, this is the first work that provides a predictive model encoding the geometric relationship between joint attention and social formation, using in-situ 3D measurements from first person cameras. This paper presents three core technical contributions: (1) a construction of a social formation feature that is scale and orientation invariant inspired by electric dipole moments; (2) a method of discovering multiple social groups using scale space extrema in a spatial distribution of social members; (3) a consolidation of social interaction data reconstructed in difference scenes, which allow us to learn and infer social saliency in a unified coordinate system. Our method can predict social saliency from a third person video or image of social interactions.

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Concept figure

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

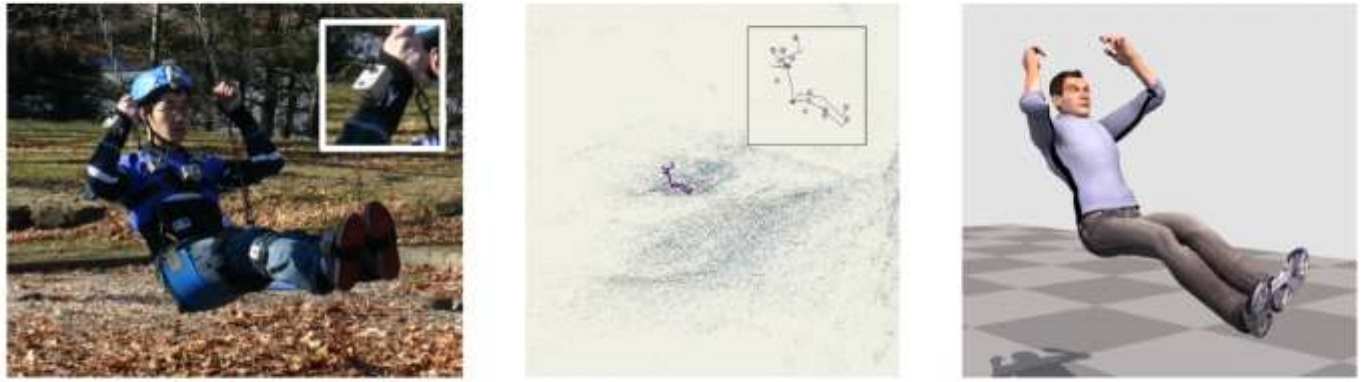
Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

Motion Capture from Body-Mounted Cameras

Takaaki Shiratori* Hyun Soo Park† Leonid Sigal* Yaser Sheikh† Jessica K. Hodgins†*
* Disney Research, Pittsburgh † Carnegie Mellon University



(a) Body-mounted cameras (b) Skeletal motion and 3D structure (c) Rendered actor

Figure 1: *Capturing both relative and global motion in natural environments using cameras mounted on the body.*

Abstract

Motion capture technology generally requires that recordings be performed in a laboratory or closed stage setting with controlled lighting. This restriction precludes the capture of motions that require an outdoor setting or the traversal of large areas. In this paper, we present the theory and practice of using body-mounted cameras

1 Introduction

Motion capture has been used to provide much of the character motion in several recent theatrical releases. In *Avatar*, motion capture was used to animate characters riding on direhorses and flying on the back of mountain banshees [Duncan 2010]. To capture realistic

Introduction Formula

1. Abstract
2. Introduction

First paragraph:

- **Motivation and problem identification**
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph

- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph

- Problem definition: input / output
- Brief presentation of results

Forth paragraph

- Contribution

1 Introduction

Motion capture has been used to provide much of the character motion in several recent theatrical releases. In *Avatar*, motion capture was used to animate characters riding on direhorses and flying on the back of mountain banshees [Duncan 2010]. To capture realistic motion for such scenes, the actors rode horses and robotic mockups in an expansive motion capture studio requiring a large number of cameras. Coverage and lighting problems often prevent directors from capturing motion in natural settings or in other large environments. Inertial systems, such as the one described by Vlasic and colleagues [2007], allow capture to occur in outdoor spaces but are designed to recover only the *relative* motion of the joints, not the global root motion.

In this paper, we present a wearable system of outward-looking cameras that allow the reconstruction of the relative and the global motion of an actor outside of a laboratory or closed stage. The cameras can be mounted on casual clothing (Figure 1(a)), are easily mounted and removed using Velcro attachments, and are lightweight enough to allow unimpeded movement. Structure-from-motion (SfM) is used to estimate the pose of the cameras throughout the capture. The estimated camera movements from a range-of-motion sequence are used to automatically build a skeleton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in tradi-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- **Limitation of existing solution**
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

1 Introduction

Motion capture has been used to provide much of the character motion in several recent theatrical releases. In *Avatar*, motion capture was used to animate characters riding on direhorses and flying on the back of mountain banshees [Duncan 2010]. To capture realistic motion for such scenes, the actors rode horses and robotic mockups in an expansive motion capture studio requiring a large number of cameras. Coverage and lighting problems often prevent directors from capturing motion in natural settings or in other large environments. Inertial systems, such as the one described by Vlasic and colleagues [2007], allow capture to occur in outdoor spaces but are designed to recover only the *relative* motion of the joints, not the global root motion.

In this paper, we present a wearable system of outward-looking cameras that allow the reconstruction of the relative and the global motion of an actor outside of a laboratory or closed stage. The cameras can be mounted on casual clothing (Figure 1(a)), are easily mounted and removed using Velcro attachments, and are lightweight enough to allow unimpeded movement. Structure-from-motion (SfM) is used to estimate the pose of the cameras throughout the capture. The estimated camera movements from a range-of-motion sequence are used to automatically build a skeleton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in tradi-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- **In this paper, …**
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

1 Introduction

Motion capture has been used to provide much of the character motion in several recent theatrical releases. In *Avatar*, motion capture was used to animate characters riding on direhorses and flying on the back of mountain banshees [Duncan 2010]. To capture realistic motion for such scenes, the actors rode horses and robotic mockups in an expansive motion capture studio requiring a large number of cameras. Coverage and lighting problems often prevent directors from capturing motion in natural settings or in other large environments. Inertial systems, such as the one described by Vlasic and colleagues [2007], allow capture to occur in outdoor spaces but are designed to recover only the *relative* motion of the joints, not the global root motion.

In this paper, we present a wearable system of outward-looking cameras that allow the reconstruction of the relative and the global motion of an actor outside of a laboratory or closed stage. The cameras can be mounted on casual clothing (Figure 1(a)), are easily mounted and removed using Velcro attachments, and are lightweight enough to allow unimpeded movement. Structure-from-motion (SfM) is used to estimate the pose of the cameras throughout the capture. The estimated camera movements from a range-of-motion sequence are used to automatically build a skeleton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in tradi-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- **Problem definition: input / output**
- Brief presentation of results

Forth paragraph
- Contribution

1 Introduction

Motion capture has been used to provide much of the character motion in several recent theatrical releases. In *Avatar*, motion capture was used to animate characters riding on direhorses and flying on the back of mountain banshees [Duncan 2010]. To capture realistic motion for such scenes, the actors rode horses and robotic mockups in an expansive motion capture studio requiring a large number of cameras. Coverage and lighting problems often prevent directors from capturing motion in natural settings or in other large environments. Inertial systems, such as the one described by Vlasic and colleagues [2007], allow capture to occur in outdoor spaces but are designed to recover only the *relative* motion of the joints, not the global root motion.

In this paper, we present a wearable system of outward-looking cameras that allow the reconstruction of the relative and the global motion of an actor outside of a laboratory or closed stage. The cameras can be mounted on casual clothing (Figure 1(a)), are easily mounted and removed using Velcro attachments, and are lightweight enough to allow unimpeded movement. Structure-from-motion (SfM) is used to estimate the pose of the cameras throughout the capture. The estimated camera movements from a range-of-motion sequence are used to automatically build a skeleton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in tradi-

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- **Insight: why this would work**
- **How this solution can overcome the challenges**

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

ton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in traditional optical motion capture, because, in our system, any visually distinctive feature in the world can serve as a marker in the traditional systems. A by-product of the capture process is a sparse 3D structure of the scene. This structure is useful as a guide for defining the ground geometry and as a first sketch of the scene for 3D animators and directors. We evaluate our approach against motion capture data generated by a Vicon optical motion capture system and report a mean joint position error of 1.76 cm and a mean joint angle error of 3.01° on the full range-of-motion sequence used for skeleton estimation. Our results demonstrate that the system can reconstruct actions that are difficult to capture with traditional motion capture systems, including outdoor activities in direct sunlight, activities that are occluded by near by proximal structures, and extended indoor activities.

Our prototype is the first, to our knowledge, to employ camera sensors for motion capture by measuring the environment and to estimate the motion of a set of cameras that are related by an underlying articulated structure. Current cameras are inexpensive, have form factors that rival inertial measurement units (IMUs), and are already embedded in everyday handheld devices. Our approach will continue to benefit from consumer trends that are driving cameras to become cheaper, smaller, faster, and more pervasive. Given the expected continuation of these technological trends, we believe that systems such as the one proposed here, will become viable alternatives to traditional motion capture technologies.

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- **Brief presentation of results**

Forth paragraph
- Contribution

ton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in traditional optical motion capture, because, in our system, any visually distinctive feature in the world can serve as a marker in the traditional systems. A by-product of the capture process is a sparse 3D structure of the scene. This structure is useful as a guide for defining the ground geometry and as a first sketch of the scene for 3D animators and directors. We evaluate our approach against motion capture data generated by a Vicon optical motion capture system and report a mean joint position error of 1.76 cm and a mean joint angle error of 3.01° on the full range-of-motion sequence used for skeleton estimation. Our results demonstrate that the system can reconstruct actions that are difficult to capture with traditional motion capture systems, including outdoor activities in direct sunlight, activities that are occluded by near by proximal structures, and extended indoor activities.

Our prototype is the first, to our knowledge, to employ camera sensors for motion capture by measuring the environment and to estimate the motion of a set of cameras that are related by an underlying articulated structure. Current cameras are inexpensive, have form factors that rival inertial measurement units (IMUs), and are already embedded in everyday handheld devices. Our approach will continue to benefit from consumer trends that are driving cameras to become cheaper, smaller, faster, and more pervasive. Given the expected continuation of these technological trends, we believe that systems such as the one proposed here, will become viable alternatives to traditional motion capture technologies.

Introduction Formula

1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- **Contribution**

ton using co-occurring transformations of the limbs connecting each joint. The reconstructed cameras and skeleton (Figure 1(b)) are used as an initialization for an overall optimization to compute the root position, orientation, and joint angles while minimizing the image matching error. Reference imagery of the capture area is leveraged to reduce drift. We render the motion of a skinned character by applying the recovered skeletal motion (Figure 1(c)).

By estimating the camera poses, the global and relative motion of an actor can be captured outdoors under a wide variety of lighting conditions or in extended indoor regions without any additional equipment. We also avoid some of the missing data problems introduced by occlusions between the markers and cameras in traditional optical motion capture, because, in our system, any visually distinctive feature in the world can serve as a marker in the traditional systems. A by-product of the capture process is a sparse 3D structure of the scene. This structure is useful as a guide for defining the ground geometry and as a first sketch of the scene for 3D animators and directors. We evaluate our approach against motion capture data generated by a Vicon optical motion capture system and report a mean joint position error of 1.76 cm and a mean joint angle error of 3.01° on the full range-of-motion sequence used for skeleton estimation. Our results demonstrate that the system can reconstruct actions that are difficult to capture with traditional motion capture systems, including outdoor activities in direct sunlight, activities that are occluded by near by proximal structures, and extended indoor activities.

Our prototype is the first, to our knowledge, to employ camera sensors for motion capture by measuring the environment and to estimate the motion of a set of cameras that are related by an underlying articulated structure. Current cameras are inexpensive, have form factors that rival inertial measurement units (IMUs), and are already embedded in everyday handheld devices. Our approach will continue to benefit from consumer trends that are driving cameras to become cheaper, smaller, faster, and more pervasive. Given the expected continuation of these technological trends, we believe that systems such as the one proposed here, will become viable alternatives to traditional motion capture technologies.

Introduction Formula

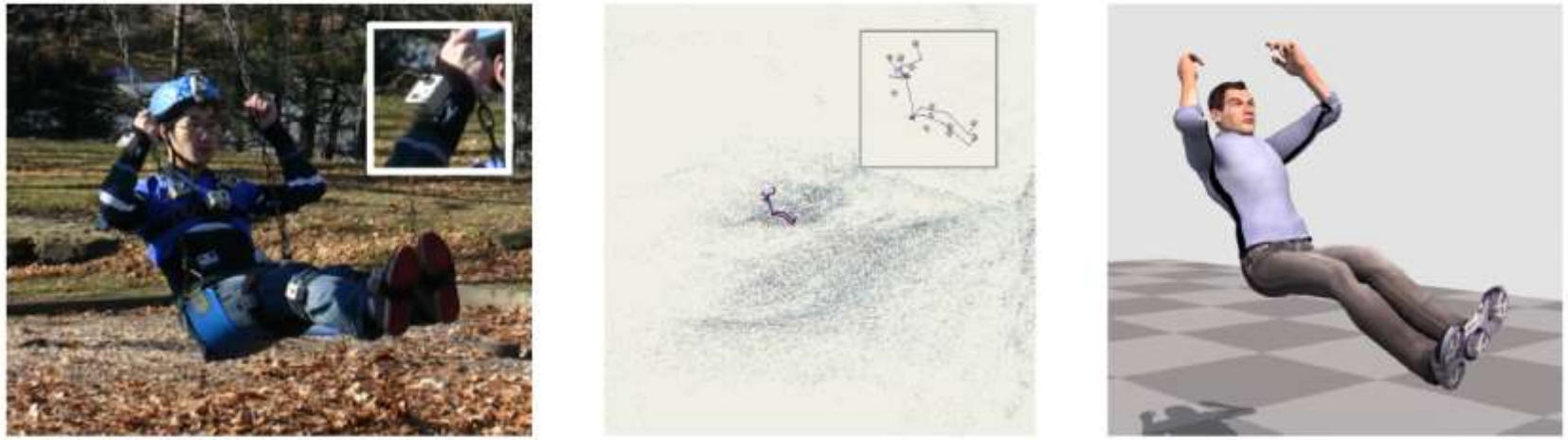
1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution



(a) Body-mounted cameras (b) Skeletal motion and 3D structure (c) Rendered actor

Figure 1: *Capturing both relative and global motion in natural environments using cameras mounted on the body.*

Introduction Formula

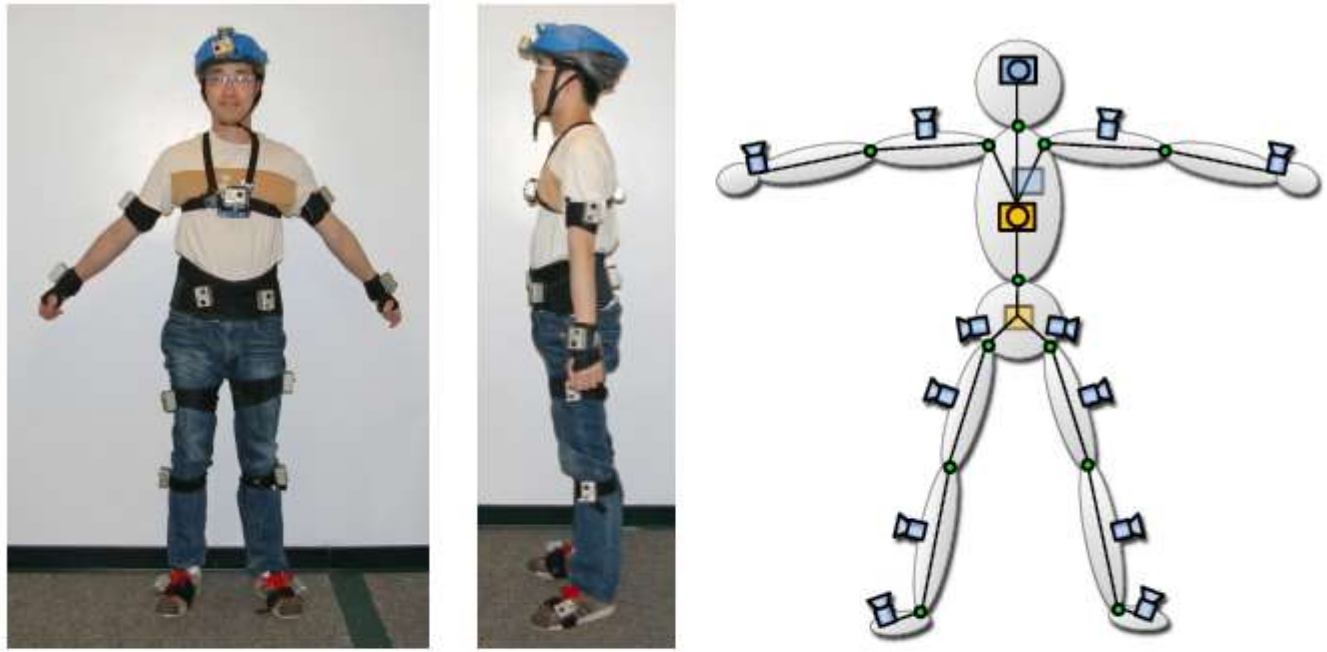
1. Abstract
2. Introduction

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution



(a) Front (b) Side (c) Skeleton

Figure 2: *Settings of cameras from (a) front view and (b) side view. (c) Illustration of skeleton and body-mounted cameras. Blue: cameras mounted on the body, and orange: cameras used as virtual cameras.*

Expected Digestable Readers

First paragraph:

- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph

- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph

- Problem definition: input / output
- Brief presentation of results

Forth paragraph

- Contribution

Expected Digestable Readers

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

Your grandma

Expected Digestable Readers

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

Your grandma

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Expected Digestable Readers

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

Your grandma

CS undergraduates

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Expected Digestable Readers

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Your grandma

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

CS undergraduates

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Reviewers

Forth paragraph
- Contribution

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Expected Digestable Readers

First paragraph:
- Motivation and problem identification
- Limitation of existing solution
- Why challenging?
- Posing research questions

Your grandma

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

CS undergraduates

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Reviewers

Forth paragraph
- Contribution

CS undergraduates

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Writing Time (I am a very very slow writer)

First paragraph:
- Motivation and problem identification 95%
- Limitation of existing solution
- Why challenging?
- Posing research questions

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

3-4 days

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Writing Time (I am a very very slow writer)

First paragraph:
- Motivation and problem identification 95%
- Limitation of existing solution
- Why challenging?
- Posing research questions

3-4 days

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

2 hours

Third paragraph
- Problem definition: input / output
- Brief presentation of results

Forth paragraph
- Contribution

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Writing Time (I am a very very slow writer)

First paragraph:
- Motivation and problem identification 95%
- Limitation of existing solution
- Why challenging?
- Posing research questions

3-4 days

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

2 hours

Third paragraph
- Problem definition: input / output
- Brief presentation of results

1 hours

Forth paragraph
- Contribution

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Writing Time (I am a very very slow writer)

First paragraph:
- Motivation and problem identification 95%
- Limitation of existing solution
- Why challenging?
- Posing research questions

3-4 days

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

2 hours

Third paragraph
- Problem definition: input / output
- Brief presentation of results

1 hours

Forth paragraph
- Contribution

0.5 hours

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

Writing Time (I am a very very slow writer)

First paragraph:
- Motivation and problem identification 95%
- Limitation of existing solution
- Why challenging?
- Posing research questions

3-4 days

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

2 hours

Third paragraph
- Problem definition: input / output
- Brief presentation of results

1 hours

Forth paragraph
- Contribution

0.5 hours

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

3-4 days

Writing Time (I am a very very slow writer)

First paragraph:
- Motivation and problem identification 95%
- Limitation of existing solution
- Why challenging?
- Posing research questions

3-4 days

Second paragraph
- In this paper, …
- Insight: why this would work
- How this solution can overcome the challenges

2 hours

Third paragraph
- Problem definition: input / output
- Brief presentation of results

1 hours

Forth paragraph
- Contribution

0.5 hours

Concept figure

Figure 1. We present a method to estimate the likelihood of joint attention called *social saliency* from a spatial distribution of social members. The inset image shows the top view of the reconstructed scene. The blue points are the points belonging to humans. The heat map shows the predicted social saliency and we overlay this map by projecting onto the ground plane in the image.

3-4 days

The three ways to write a good paper are refinement, refinement, and refinement.