1
Introdução

Esta pesquisa tem por objetivo reconstruir câmeras utilizando fotos reais e modelos computacionais de edificações por meio de uma técnica semi-automática. A reconstrução de câmera é um dos problemas fundamentais da visão computacional e consiste, em linhas gerais, no cálculo de parâmetros que compõem um modelo matemático de câmera (chamado de câmera virtual) a partir da imagem de uma cena real. Tal câmera virtual é a representação computacional de uma camera real e pode ser utilizada para, por exemplo, interpretar e reconstruir a estrutura tridimensional de uma cena real a partir de fotos ou vídeos digitais.

Na área de robótica, por exemplo, câmeras eletrônicas são instaladas em dispositivos móveis que podem se deslocar pelo mundo e, por meio da detecção e interpretação de características nas imagens capturadas, recuperar uma câmera virtual que é usada para prover orientação em relação a um determinado objeto conhecido do mundo.

As técnicas de reconstrução de câmera da visão computacional são frequentemente usadas em conjunto com técnicas de realidade virtual para dar origem a novas aplicações chamadas de aplicações de realidade aumentada [Raposo et al., 2004]. Como tais câmeras virtuais permitem sintetizar computacionalmente imagens com aparência realista, os sistema de software de realidade aumentada as utilizam para combinar elementos reais e virtuais em uma mesma foto digital. É esse tipo de procedimento que é usado, por exemplo, em partidas televisionadas de jogos de futebol quando elementos de propagandas virtuais são inseridos em meio a imagens do campo como se fizessem parte da cena real.

Aplicações desse tipo são usadas também em outras atividades em que elementos computacionais são inseridos em imagens reais para auxiliar algum tipo de procedimento humano, tais como a de montagem e inspeção de equipamentos de engenharia, auxílio a procedimentos médicos, em sistemas de arquitetura, dentre outras.

Existem também aplicações de realidade aumentada desenvolvidas especificamente para prover auxílio a visitas virtuais a monumentos, objetos

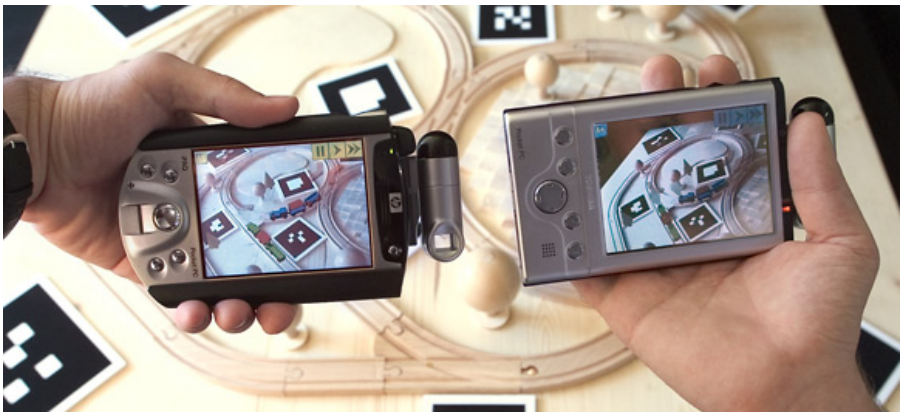

Figura 1.1: Aplicação de realidade aumentada. Marcadores planares são colocados sobre a mesa e utilizados para reconstruir câmeras virtuais que permitem inserir objetos sintéticos nas imagens capturadas pelas câmeras dos dispositivos móveis, como as locomotivas no exemplo.

históricos e a museus [Braga, 2007]. Em tais visitas *aumentadas*[1] a edificações antigas, freqüentemente em ruínas como as da Grécia antiga, turistas utilizam dispositivos especiais como visores[2] de imersão, sensores e computadores portáteis para visualizar partes reais dessas ruínas integradas com partes sintéticas de modelos computacionais. Isso permite que eles enxerguem como essas edificações eram em suas formas originais. Um exemplo desse tipo de aplicação pode ser encontrado no projeto Archeoguide [Vlahakis et al., 2002].

É importante, no entanto, distinguir dois tipos de aplicações de realidade aumentada: as que demandam um tempo de resposta mínimo e utilizam um paradigma de funcionamento de **tempo real** e as **não interativas**[3]. Nas aplicações de tempo real, as imagens de entrada são capturadas ao vivo e processadas em um curto espaço de tempo para que seja possível produzir informações que podem ser acessadas de modo interativo. Esse paradigma de funcionamento requer grande eficiência de processamento das imagens porque é preciso que se mantenha uma interatividade constante com o usuário mantendo sempre o fluxo de processamento das novas imagens capturadas. Em outras palavras, tais aplicações capturam e interpretam cada *frame* do fluxo de imagens capturadas e produzem resultados imediatos para cada um deles,

[1] O termo "aumentada" indica que uma determinada ação ou procedimento é auxiliado por uma aplicação de realidade aumentada.

[2] Visores de imersão (*Head-mounted display*) são dispositivos especiais que funcionam como óculos que podem inserir imagens geradas por computador. Dessa forma um usuário pode ver imagens reais de forma integrada com imagens sintéticas.

[3] O termo "não interativo" é usado aqui para designar aplicações cujo tempo de resposta não permite a interação com cenas capturadas ao vivo, como nas chamadas aplicações de tempo real. No entanto, tal nomenclatura não implica que essas aplicações sejam obrigatoriamente não dependentes da interação do usuário.

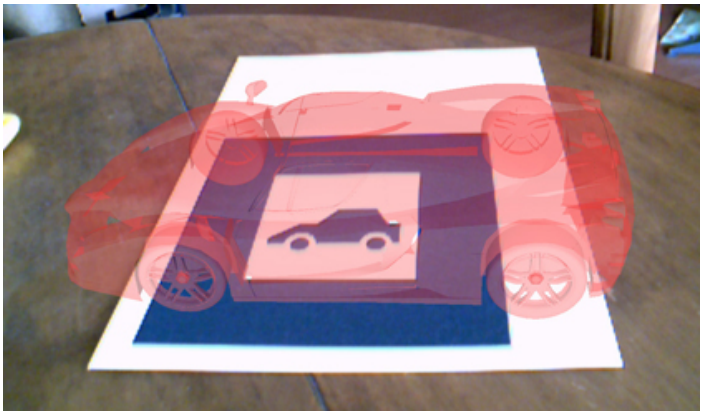

Figura 1.2: Aplicação de realidade aumentada. O padrão planar mostrado impresso em folha de papel possui uma geometria conhecida. A partir dele é possível reconstruir uma câmera virtual para sintetizar objeto mostrado, no caso, um automóvel vermelho. O procedimento foi feito utilizando a biblioteca Artoolkit.

sempre um tempo mínimo de resposta. Já nas aplicações que não demandam tempo real, o processamento é feito em cima de imagens ou vídeos obtidos previamente, ou seja, que não são adquiridos em tempo real. Nesses tipos de aplicação não há a necessidade de um tempo de resposta tão curto, pois o conjunto de dados de entrada é limitado e fixo.

Porém, tanto as aplicações de tempo real quanto as não interativas deparam-se com o problema da reconstrução de câmera, mas utilizam estratégias distintas para resolvê-lo, que se adaptam às suas propostas de funcionamento.

Para reconstruir uma câmera virtual é preciso, em muitos casos, desenvolver algoritmos que localizem características na imagem que correspondam a características conhecidas do mundo, o que representa um dos principais problemas da visão computacional [Fischer and Bolles, 1981]. Existem, contudo, diversas estratégias para resolvê-lo. Algumas aplicações utilizam padrões de calibração cuja geometria é conhecida e que podem ser facilmente detectados na imagem, como os mostrados nas figuras 1.2 e 1.1. Outras usam marcadores eletrônicos que fornecem orientação e posição em relação a um ponto adotado como origem, como dispositivos de radio frequência, laser ou GPS [Chou, 2000]. Algumas, ainda, não usam marcadores de nenhum tipo e utilizam aprendizado de máquina para identificar a estrutura da cena e interpretá-la [Saxena et al., 2007; Delage et al., 2007].

Em aplicações que não usam marcadores eletrônicos, a maior dificuldade está em encontrar um conjunto de características da imagem que têm relação com o mundo. Ou seja, localizar pontos ou segmentos na imagem e inter-

Figura 1.3: Exemplo de realidade aumentada aplicada a ruínas de edificações. Na imagem acima uma câmera virtual foi reconstruída a partir de marcadores eletrônicos para inserir o modelo do templo de Zeus, em Olympia, em sua aparência original. Projeto Archeoguide.

pretá-los como sendo partes conhecidas da cena. Através desse conjunto de características relacionadas, torna-se possível aplicar diferentes métodos para efetuar a reconstrução de câmera. Sendo assim, o principal problema a ser resolvido passa a ser como encontrar tais características relacionadas. Dependendo do tipo da aplicação, tempo real ou não interativa, pode-se utilizar estratégias diferentes.

Aplicações de tempo real freqüentemente utilizam câmeras instrumentadas e marcadores eletrônicos como GPS, pois precisam ser eficientes e automatizadas. Tais dispositivos permitem que a câmera seja recuperada de forma rápida e sem que seja necessário realizar processamento de imagens, o que demanda um esforço computacional alto. Porém, aplicações que não têm compromisso com interação em tempo real, podem lançar mão de estratégias mais simples que dependem inclusive de marcações nas imagens fornecidas por um usuário.

A abordagem adotada neste trabalho consiste justamente em recuperar câmeras a partir de uma única foto que não é obtida em tempo real, sem a utilização de marcadores especiais e com possível auxílio do usuário.

Por não haver marcadores aplicados à cena, o próprio modelo retratado nas imagens é usado como padrão de calibração. Isso faz com que esse modelo assuma, então, o papel de marcador, mas também exige que se conheça suas proporções reais e que as imagens de entrada sejam processadas para que relações entre elas e esse modelo sejam identificadas. O tema aqui tratado tem como foco principal o relacionamento entre imagens reais e modelos virtuais de

Figura 1.4: Exemplo de aplicação de realidade aumentada não interativa que utiliza marcações feitas pelo usuário para recuperar, dentre outras coisas, a câmera que gerou a imagem. Na imagem acima, as arestas em verde são marcações feitas por um usuário. A imagem foi extraída do trabalho feito por Paul E. Debevec em 1996 em sua tese de doutorado.

edificações, através de um casamento direto entre eles, utilizando uma única foto por vez e uma técnica semi-automática de reconstrução de câmera.

O caso de estudo realizado por esta pesquisa envolve principalmente um conjunto de fotos e um modelo aproximado de uma edificação antiga. Contudo, como será apresentado, o método proposto não impõe restrições em relação ao tipo de modelo tratado, de modo que é possível utilizar também modelos de objetos cuja estrutura é conhecida com mais precisão, como modelos de engenharia ou modelos CAD de edificações modernas.

1.0.1
Motivação e Objetivo

A motivação para o tema surgiu no decorrer de um projeto acadêmico executado pelo grupo de Realidade Virtual do Tecgraf Puc-Rio. Desejava-se realizar o casamento entre fotos e modelos tridimensionais e, dessa forma, construir uma espécie de visita virtual em que se pudesse observar ora a imagem da ruína em seu estado presente (como a foto a representa atualmente) e ora o modelo sintético com a aparência original da ruína. Algumas dificuldades iniciais foram identificadas em relação a esse problema:

1. As imagens reais da edificação continham uma grande quantidade de ruído e elementos como vegetação densa e rachaduras que dificultavam a localização de um conjunto de correspondências com o modelo virtual;

2. O modelo da edificação em sua forma original foi criado manualmente a partir de documentos históricos, plantas antigas e fotos, o que faz com que sua geometria seja possivelmente imprecisa.

Pelo motivos expostos, foi preciso criar uma técnica capaz de localizar características boas em fotos ruidosas de ambientes externos e que fosse flexível o suficiente para se adaptar da melhor forma possível aos modelos possivelmente imprecisos. O tipo de característica adotada para uso nas imagens foi o segmento de reta, ao invés do ponto (quinas), pois esse se mostrou mais robusto quando aplicado à fotos de edificações. Assim, desenvolveu-se um método simples que possibilita o casamento entre modelos e fotos de edificações — tipicamente ruínas — com o mínimo de erro possível, ou seja, que resulta em um ajuste visualmente aceitável.

Um outro objetivo a ser alcançado ao reconstruir as câmeras, é obter uma espécie de registro tridimensional das fotos que elas representam, possibilitando então uma navegação espacial entre essas fotos como em um álbum tridimensional. Tal navegação permite, por exemplo, que a partir de uma dada foto (em uma posição de câmera reconstruída), se navegue para a foto que está mais próxima usando alguma direção específica (por exemplo, a foto mais próxima à direita).

Assim, os principais requisitos funcionais deste projeto são:

- Reconstrução e posicionamento de camera virtual utilizando um conjunto esparso de fotos, modelos computacionais e marcações auxiliares feitas pelo usuário;
- Capacidade de registrar as diversas câmeras reconstruídas;
- Navegação tridimensional entre câmeras registradas.

Tendo em vista as limitações mencionadas, optou-se por uma abordagem que consiste em utilizar uma técnica semi-automática de reconstrução e ajuste de câmera. O termo "semi-automática" significa que o usuário tem um papel importante para que a recuperação da câmera e o conseqüente casamento entre a foto e modelo aconteçam corretamente: ele fornece marcações de boa qualidade na foto (com auxílio do computador) que serão usadas para localizar características do mundo e recuperar a câmera. Tal intervenção manual se mostra necessária devido a complexidade em se localizar e interpretar características bem definidas nas fotos de forma totalmente automática.

Em casos simples, no entanto, — em que as fotos de entrada possuem pouco ruído — a técnica desenvolvida permite que o usuário manipule *apenas a imagem* para que as associações com pontos do mundo sejam detectadas, de tal modo que ele não precise criar diretamente associações entre segmentos da imagem e do modelo (Seção 3.6). Nestes casos, as associações são calculadas automaticamente usando-se apenas uma posição e orientação iniciais de câmera fornecidas pelo usuário, como será mostrado. Contudo, a técnica de reconstrução de câmera utilizada (Seção 2.2) pode funcionar também com um conjunto de associações imagem-modelo fornecidas explicitamente pelo usuário, o que permite que a reconstrução ocorra mesmo em casos complexos.

Dessa forma, o método proposto tem como objetivo reconstruir câmeras através de técnicas integradas que fornecem mecanismos que ajudam o usuário a criar um conjunto de correspondências entre modelo e imagem (fundamentais para a calibração e o posicionamento), de tal modo que ele tenha o mínimo de esforço possível para fazê-lo. A partir dessas associações criadas de forma assistida pelo usuário, a câmera é recuperada para possibilitar a navegação entre as todas as imagens fornecidas como entrada. A técnica deve prover o melhor posicionamento e reconstrução de câmera possível entre as fotos e o modelo virtual.

Por se tratar de uma técnica semi-automática e demandar interação com o usuário, há certos requisitos não funcionais que devem ser respeitados:

- Eficiência: executar em um intervalo de tempo minimamente aceitável do ponto de vista do usuário, prezando boa experiência de uso;
- Simplicidade: ser pouco trabalhoso e prover boa produtividade.

Em suma, o objetivo da dissertação é desenvolver um método semi-automático e eficiente para casar imagens digitais de edificações com seus modelos computacionais tridimensionais, sem uso de marcadores aplicados ao objeto real, obtendo como produto desse casamento o modelo computacional da câmera que originou a foto real. Esta técnica deve ser simples e demandar pouca interação do usuário.

Os principais pontos desenvolvidos por esta dissertação são:

- Uma técnica automática de calibração utilizando meta-dados EXIF, extraídos de arquivos de dados de imagens;
- Uma técnica semi-automática de calibração utilizando pontos de fuga;
- Um método para extrair arestas estruturais de um modelo CAD qualquer que facilita o processo de localização de correspondências entre modelo e imagem;

- Um método semi-automático para calcular correspondências entre modelo e imagem onde o próprio modelo é utilizado para restringir a área de busca na imagem;
- Uma técnica de posicionamento de câmera que utiliza correspondências entre segmentos de reta em uma imagem e segmentos de um modelo geométrico tridimensional;
- Uma aplicação de exemplo que implementa todos os conceitos propostos nesta dissertação e provê solução completa para recuperação e registro de câmeras. A aplicação possibilita efetuar o casamento semi-automático entre modelos CAD e fotos, além de disponibilizar ferramentas para compará-los e prover uma nova experiência de navegação tridimensional entre as diversas fotos registradas.

Esta dissertação está organizada em 6 capítulos, incluindo esta introdução que constitui o primeiro deles. O Capítulo 2 expõe de forma resumida os principais assuntos que estão relacionados diretamente com essa pesquisa, provendo uma linha de raciocínio que introduz os conceitos que serão apresentados ao longo desta dissertação. No Capítulo 3, a técnica para calibração e posicionamento de câmera desenvolvida aqui é exposta em detalhes. O Capítulo 4 apresenta a aplicação desenvolvida como produto desta pesquisa para ilustrar o método proposto e as diversas técnicas que o compõem. No Capítulo 5 são expostos resultados visuais e quantitativos e, finalmente, o Capítulo 6 conclui esta pesquisa e apresenta propostas para trabalhos futuros.